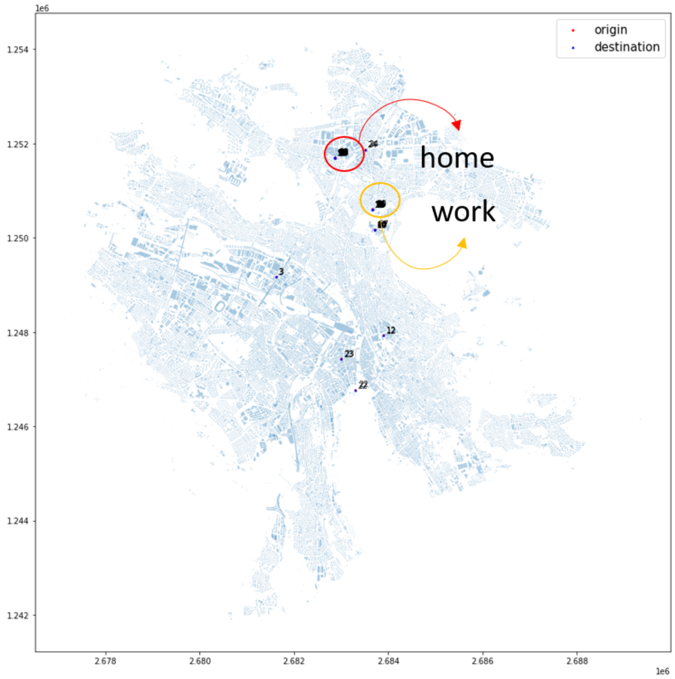


Understanding Public Transport Commuting Trips from GPS Tracking Data

Mariana A. Costa

Alessio D. Marra

Francesco Corman

STRC conference paper 2022 **May 10, 2022**

STRC | **22nd Swiss Transport Research Conference**
Monte Verità / Ascona, May 18-20, 2022

Understanding Public Transport Commuting Trips from GPS Tracking Data

Mariana A. Costa
IVT
ETH Zürich
CH-8093 Zurich
mariana.de-almeida-costa@ivt.baug.ethz.ch

Alessio D. Marra
IVT
ETH Zürich

Francesco Corman
IVT
ETH Zürich

May 10, 2022

Abstract

Understanding the choices of passengers and preferences in public transport has been traditionally done through surveys. Global Positioning Systems (GPS) tracking data is a low-cost and efficient alternative to those surveys. New technologies allow automatic collection of data that significantly reduce the burden placed on users, with low battery usage, satisfactory spatio-temporal precision and restricted (or absent) user interaction in the form of manual inputs. Passive tracking can capture the long-term behaviour of passengers' trips, but with such unprecedented levels of data, it is pivotal to extract relevant information that ensures a dynamic response from the service providers, especially in multimodal networks, where the route options of users can be infinitely many. A significant part of passengers' trips are commuting trips from home to work and vice-versa, therefore, understanding passengers' behaviour in commuting trips is a key factor to enhance urban transport planning. This paper uses unlabelled GPS tracking data collected by a smartphone application (*ETH-IVT Travel Diary*) to explore strategies for unsupervised classification of work and home activities and, based on the imputed labels, understand behavioural aspects of passengers on commuting trips. Disparities in the recurrence of routes and times are studied, suggesting important differences between non-commuting and commuting trips, as well as home-work and work-home trips. Passengers' willingness to walk and to make transfers and the impact of disruptions are also investigated based on the perspective of recurrent behaviour in commuting trips.

Keywords

Public transport, Route choice, Passengers behaviour, GPS tracking data, AVL data

Contents

List of Tables 1

List of Figures 1

1 Introduction 3

2 State of the art 6

3 Unsupervised Classification of Activities 10

4 Behavioural Aspects of Commuting Trips 21

4.1 Mode share in commuting trips *vs* non-commuting trips 24

4.2 Recurrence of modes and routes in commuting trips 25

4.3 Comparison of commuting trip choices with timetable and realized times 37

4.4 Impact of disruptions on commuting trips routes and information available to the traveller 41

5 Conclusions and Future Research 46

6 References 48

List of Tables

1 Comparison between clustering based on distance parameters only and clustering based on distance and time parameters for different values of DBSCAN ϵ parameter. 14

2 Comparison between clustering based on distance parameters only and clustering based on distance and time parameters for different strategies. 19

3 Sociodemographic and behavioural characteristics of travellers in the *ETH-IVT Travel Diary* survey 20

4 Summary of activities according to chosen clustering strategy. 22

5 Statistics for the proportion of similar trips across all users in commuting trips. 36

List of Figures

1 Clusters for the same user based on DBSCAN with two values of the eps parameter: 500m (left) and 100m (right). 13
2 % of agreement with distance-only strategy 17
3 Mode share for the *ETH-IVT Travel Diary* dataset: all trips, non-commuting trips and commuting trips (home-work and work-home) 24
4 Histogram of distance in minutes from average departure times 26
5 Mode Share of Preferred (PT) Mode of Transport for each Individual in Commuting Trips 27
6 Counts of Frequencies of the Preferred (PT) Mode of Transport per Trip Category 29
7 Counts of Frequencies of the Preferred (PT) Line per Trip Category 31
8 Box Plots: Breakdown of Trip Times (Home-Work *vs.* Work-Home) 33
9 Similarity Metrics for Users with Home-Work and Work-Home Trips 35
10 Comparison of commuting trips path choices made by travellers with ordered timetable CS positions 38
11 Comparison of commuting trips path choices made by travellers with sorted actual CS positions 39
12 Box Plots of proportions (for each traveller) of trips in the top 1 or in the top 10 of CS over total number of trips (home-work and work-home trips) 40
13 Box Plots of the difference in time duration (in minutes) of timetable and actual routes 43
14 Commuting trips outside of CS: depiction of factors. 45

1 Introduction

Identifying commuting patterns is a key element for enhancing urban networks, as they reflect the long-term behaviour of individuals and have a considerable impact on human mobility (Kung *et al.*, 2014). In particular, efficient and reliable Public Transport (PT) is an effective tool to mitigate traffic congestion and alleviate emissions, while significantly reducing car dependency (Ma *et al.*, 2017) and the pressure on scarce road space. While high-quality PT links to employment centres are shown to encourage switches away from car commuting (Clark *et al.*, 2016), to stay an attractive alternative to the population, PT planners and operators must identify opportunities and keep up to speed with transit commuting behaviour on an individual level. PT behavioural studies have been traditionally done through surveys, either stated preference (SP) or revealed preference (RP). In terms of intermodal travel, for which many possible alternatives of routes exist, these methods based on surveys may produce biased results: in SP because the user is not experiencing the trip himself (and daily mobility is usually inferred based on a reference day) and, in RP, because the path sampling process is very complex, causing the few surveys reported to be useful only for long-distance travel situations as highlighted by de Freitas *et al.* (2019). Nevertheless, these surveys are costly, time-consuming and frequently result in low sampling rates (Ma *et al.*, 2017).

Recent developments in technologies to acquire data, as well as emerging statistical methods for the analysis of such data, allow better investigation of the individual commuting behaviour, which is crucial for the purpose of achieving the full potential of intelligent transportation systems, a challenge that is central to creating environmentally sustainable cities. Global Positioning Systems (GPS) tracking data is a low-cost and efficient alternative to traditionally used travel surveys. New technologies allow collecting such data automatically, for instance through smartphone applications that significantly reduce the burden placed on users, with low battery usage, satisfactory spatio-temporal precision and restricted (or absent) user interaction in the form of manual inputs (Marra *et al.*, 2019). In times of unprecedented data availability and global competitiveness, it is pivotal to extract relevant information to ensure an in-depth understanding of the passengers' behaviour and dynamic response from the service providers. In this sense, a key element in understanding the route choices of users is to know which alternatives are available to them as well as their behavioural characteristics.

Multimodal networks are a reality in many big cities worldwide. The route options of users can be infinitely many, as different transport modes options may be available within the same path. Travelling in such an intermodal manner can help mitigate the effects of a potentially saturated transport infrastructure, however, it poses an additional burden both on the users, which will likely have to deal with transfers and the uncertainty associated with them, and on the urban traffic planners, as route and mode choice behaviour of travellers have to be anticipated. In fact, these approaches have been traditionally modelled separately, with route choice models tending to focus on unimodal paths, and mode choice models not modelling the route in detail (de Freitas *et al.*, 2019). Nevertheless, intermodal travel requires that both route and mode are considered simultaneously, so that feasible alternatives can be extracted from the multimodal network. For simplicity, the term route choice will be used herein, where it is assumed that the mode choice is implicit in the route chosen by the traveller. Another assumption is that intermodal trips are considered in their broader definition, so any trip involving at least one mode of transport is considered as a trip.

The willingness to conduct intermodal trips is dependent on the travellers' socioeconomic characteristics. Intermodal travel patterns are mostly observed when a transit subscription allows free at the point-of-use public transportation (de Freitas *et al.*, 2019). Previous research (Marra *et al.*, 2019) has indicated that most users will display similar behaviour when making decisions on the routes, for instance, by prioritizing routes with shorter times and a smaller number of connections. Equally important is the task of distinguishing the traveller's behaviour in the case of service disruptions/disturbances and also according to the level of disruption. For multimodal networks, a disruption is better defined from the operational perspective, taking into account delays or missed trips, not only failed links for a certain amount of time, which is usually the definition utilised for railway/metro networks (Marra and Corman, 2020b). For instance, in the case of small disruptions, although most users may opt for not changing their original plans (Marra *et al.*, 2019), investigating why some users choose an alternative route may reveal some important behavioural characteristics, such as willingness to walk or willingness to accept a potential delay. Such travellers' attitudes and reactions to transport are usually not present in surveys or microcensus, instead, they must be inferred from observed data. In the literature, attempts to draw a generic profile for intermodal travellers based on the individual's socioeconomic and socio-demographic characteristics acquired through questionnaires and microcensus are abundant. de Freitas *et al.* (2019) summarize some relevant works. However, although these studies with categorical data exist, they do not exactly reveal the behaviour of a random user confronted with many possible routes and mode choices. For example, if a traveller takes the same train to work every day, but a ten-minute delay on a certain day makes the traveller opt for an alternative route, then the 'willingness' to change the route, even at minimal disturbances, can be considered high. On the contrary, if the traveller did not opt for an alternative route, it could mean two things: information on the delay was not available to the traveller and/or the traveller decided to stick with the route even though an alternative route would have led to inferior travel time.

This paper uses travel diaries collected by a smartphone application called *ETH-IVT Travel Diary* consisting of 2901 public transport trips of 172 users in the city of Zürich (Switzerland). The application allowed (continuous) passive tracking, and activities, trips and modes were identified through a mode detection algorithm, as described in Marra *et al.* (2019). In addition, the algorithm could identify the public transport line and vehicle used, by appropriate matching with Automatic Vehicle Location data (AVL) of the Zürich public transport network. Hence, for each public transport stage the following information was detected: the mode (bus, tram or train), the line, the specific vehicle of that line, the user's departure stop and time, the user's arrival stop and time. Next, a choice set (CS) generation algorithm was proposed and tested for this dataset, as detailed in Marra and Corman (2020a), to identify the available alternatives to users according to the timetable and also according to actual (realized) times. A Mixed Path Size Logit model (an extension of the Multinomial Logit) was used to test the CS algorithm, showing an accuracy of 94%, i.e. 94% of the times the CS contained the same alternative chosen by the user, a very high coverage of 2734 trips out of 2909. The good fit of the Mixed Path Size Logit suggested that a linear utility function was sufficient to explain most of the observed route choices. Only about 6%, or 175 trips, did not have a match within the first 100 CS trip options.

This paper extends the results in Marra *et al.* (2019); Marra and Corman (2020a) by further investigating the GPS tracking data collected, the paths generated by the mode detection algorithm and also the CS generation algorithm, to determine patterns and logical reasoning to i) distinguish among commuting (trips from home to work, and vice-versa) and non-commuting trips; ii) summarize the main behavioural aspects and characteristics of travellers in commuting trips and, whenever applicable, compare the differences with non-commuting trips; iii) assess the impacts of disruptions in commuting trips and infer whether online information was available to travellers and; iv) investigate path choices not identified in the CS. This study is motivated by existing literature (e.g. Lima *et al.*, 2016; Levinson and Zhu, 2013) supporting that mobility patterns and route choice behaviour are regular for commuting trips, suggesting that travellers usually opt for the same few alternatives. Under this hypothesis, it could be assumed that travellers make conscious and recurrent path choices for their commuting trips, and collecting statistics on these trips would be a key element for service planners to provide a better service. In addition, changes in a common behaviour could reveal important users' characteristics under disruptions.

In the case of Zürich, some characteristics of the network may diminish the impacts of possible service disruptions. For example, Marra and Corman (2020b) show that the frequency of service can compensate for delays or single failures, making individuals' willingness to opt for alternative paths to be small. In theory, in a network with high reliability, where users commonly opt for the same path choices, the low variance inherent in these commuting trips could make it difficult for traditional path-based algorithms to generate distinct route alternatives, which, in turn, would lead to biased parameters estimates in the route choice models. One of the goals of this work is to investigate, in the highly reliable network of Zürich, how recurrence patterns are observed for travellers that rely on public transport for their commuting trips.

2 State of the art

Most previous works on commuting patterns focus on private vehicle traffic (including ride-sharing), for which data is more accessible (Zhao *et al.*, 2019). For example, Zhao *et al.* (2019); Hong *et al.* (2020) use automatic vehicle identification (AVI) and automatic license plate recognition (ALPR) data, respectively, to investigate the commuting behaviour of private vehicle travel in different cities in China. Mobile phone data, such as call detail records (CDRs), is also reported as a comprehensive and versatile data source for studying large-scale human mobility (see, e.g., Kung *et al.* (2014)), although it is usually "car-heavy".

In terms of PT, automated fare collection (AFC) data and GPS records are amongst the most reported data types for studying commuting behaviour, although, as highlighted by Ma *et al.* (2017), many works on transit behavioural studies define commuting only in terms of repeatability of temporal activities (e.g. users travelling four days or more per week are considered commuters). Only a few works model commuting behaviour both in terms of spatial and temporal regularity. Ma *et al.* (2013) use DBSCAN, a density-based clustering algorithm, to investigate spatial and temporal travel patterns in Beijing, and then utilise a K-means++ algorithm and a Rough Set based approach to measure travel regularity. Still investigating travel patterns in Beijing, Ma *et al.* (2017) propose a series of data mining methods using smartcard data, and find out that the majority of commuters depart around morning peak hours (7:00-9:00) and return during evening peak hours (17:00-19:00), whereas a clear pattern is not observed for noncommuters. Commuters also have an associated high number of travelling days, with a mode of 21 days (approximately the number of weekdays in a typical month), whereas 90% of noncommuters travel below 10 days. Lastly, they conclude that when the distance between residence and workplace is far, commuters are less likely to opt for PT.

Goulet-Langlois *et al.* (2016) use smartcard data combined with socio-demographic information to identify clusters of users with similar activity sequence structures. Their main finding related to commuting behaviour is that, while conventional working days are an important element of structure for many passengers, they do not structure the activity sequence of over 40% of frequent users, which could have implications on the usual 'typical commuter' modelling profile. Ortega-Tong (2013) also explores the similarity in travel patterns from riders with smartcards combined with socio-demographic characteristics to identify clusters with similar structures. Two clusters are identified as being composed of commuters, with the first corresponding to a group mostly composed of students and, the second, of workers. Two important distinctions among these two clusters of commuters refer to students having shorter school days and their preference for bus trips due to lower fares.

Zhou *et al.* (2014) study commuting efficiency along with the bus network in the Beijing metropolitan area by combining smartcard and travel survey data. They use linear programming to compute the minimised mean commuting costs and calculate the excess commuting (surplus travel time from people that do not take the optimal route to work). They conclude that easy access to work is one of the key factors involved in bus commuting (over car). Kusakabe and Asakura (2014) also use smartcard data but with the aim of analysing behavioural features to classify trip purpose by utilising a naïve Bayes classifier. An interesting finding is related to changes in the commuting pattern during holiday seasons, in which commuters significantly reduce their trip frequency, thus affecting the total number of trips.

To the best of the authors' knowledge, longitudinal studies on commuting trips based solely on GPS tracking data, that also investigate the effects of possible disruptions to infer behavioural characteristics of the travellers, are yet to be published, so this paper aims to fill this research gap. While both GPS tracking and AFC enable collecting disaggregated data on passenger boarding and alighting and, thus, studying spatial and temporal regularity, AFC data has the obvious drawback of lacking the information on the actual origin and destination of the trip, so that, for instance, the exact location of residence and workplace is not known, unless this information is otherwise made available by the smartcard owner. In this sense, GPS tracking data, nowadays mostly acquired through smartphone applications (Cottrill *et al.*, 2013), is a low-cost and a more efficient alternative as it can capture all the user's movements throughout the day. While most of the available smartphone applications use a prompted recall approach, requiring the user to manually add some trip details, such as mode, transit fare and trip purpose (e.g. Cottrill *et al.*, 2013; Molloy *et al.*, 2020, 2021), the problem of manually annotating trips places a significant hurdle on data collection Marra *et al.* (2019). One of the goals of the *ETH-IVT Travel Diary* survey was to be completely based on passive GPS tracking (no user inputs on trips) and, by placing a very low burden on users, to capture long-term behaviour, thus allowing a better understanding of day-to-day variability and the user's response to potential disruptions in public transport. After acquiring the data and according to the methodology proposed by Marra *et al.* (2019), the problem of mode detection can be divided into four main tasks, as follows:

1. Data cleaning: consists of identifying erroneous, incomplete or irrelevant information or records in a database and modifying/deleting these records to obtain a consistent database. In the case of raw GPS data, filtering and smoothing are the two main techniques used for the purpose of data cleaning, where the filtering removes data that do not represent the user's real position and smoothing reduces the random noise present in the data.

2. Trip/activity identification: a user's day can be described by alternating trips and activities, where a trip (walk or ride) is defined by a sequence of points located apart from each other, indicating movements, whereas an activity is defined by a sequence of points next to each other, indicating that the user is in the same place for some period.

3. Trip segmentation: after identifying the trips, these can be segmented first into walk-stages or other-stages, where walk-stages are simply the paths that the user walked, and other-stages represent all the other means of transport (car, bus, train or other vehicles).

4. Mode detection: a mode detection algorithm matches the exact public transport vehicle used by the traveller based on GPS tracking data and AVL data. By doing so, it is possible to classify the trips into public (bus/train/tram) or private (bike/car) modes of transport.

Mode detection is a fundamental input to consider when examining the passenger's choice in a route choice problem, which aims to understand and make inferences regarding the chosen path of the passenger in a transport network. However, to understand passenger's behaviour, it is also fundamental to know the available alternatives in terms of route choices, including information regarding possible disturbances. Route choice of the passenger is usually tackled in literature from two perspectives: the choice model, which aims to determine the actual chosen route out of a small set of options (also known as the choice set, CS), and the choice set (CS) generation, which focuses on identifying all the relevant alternatives to passengers (Bovy *et al.*, 2008). Regarding the latter, finding all alternatives available to the users may be computationally prohibitive, whereas restricting the size of the CS may affect the coverage precision. Moreover, defining the relevant alternatives, or paths, is not a straightforward task, as it involves an attempt to model an assumed behaviour. Most works assume that passengers behave rationally, and always seek to minimize a certain cost function or, alternatively, to maximize a certain utility function, which can depend on network attributes and possibly socioeconomic characteristics of passengers (Zimmermann and Frejinger, 2020). A good model is one that can properly identify a cost/utility function from a set of observed trajectories.

CS generation algorithms can be deterministic or stochastic. Among the deterministic, many works identify minimum cost paths, consisting of the shortest paths using different cost functions. Some other works constrain the generated CS by assuming the user chooses routes based on some factors, such as the number of transfers (Zimmermann and Frejinger, 2020). The stochastic CS algorithms include a stochastic factor in the generation of each path. For this paper, the CS generation algorithm applied to the *ETH-IVT Travel Diary* dataset and described in the work of Marra and Corman (2020a) is used. The model assumes that passengers are rational and will always choose the route that maximizes their utility, which is defined as a function of travel components, namely transfer times, walking times and times on PT (bus, tram and train), and also adds a penalty for the number of transfers and a correction term for overlapping paths (the Path Size cost). The proposed algorithm is evaluated on another large-scale (labelled) tracking dataset and achieves high precision both in terms of coverage (over 94%) and model estimation (in terms of high R^2).

The next sections assume that the mode detection algorithm is successfully able to identify the correct paths taken by each user in the *ETH-IVT Travel Diary* dataset. Moreover, the path taken by the user is contrasted with the other PT alternatives available, as identified by two CSs generated with different information provision: a CS assuming the PT timetable (planned) times and a CS assuming the actual (realized) times for the PT modes available according to the AVL data available. For more details on both the mode detection and CS generation algorithm, the reader is referred to Marra *et al.* (2019); Marra and Corman (2020a).

3 Unsupervised Classification of Activities

A crucial aspect of the *ETH-IVT Travel Diary* application (Marra *et al.*, 2019) was related to reducing the burden placed on users. An evident downside was that labelled data (such as user confirmation of modes of transport and activities for each trip) was not available. Although information of routes and the relevant modes of transport can be obtained by appropriately matching GPS time and position data with real-time public transport information (e.g. AVL data), classifying activities using only spatio-temporal data is challenging. Even some traditional unsupervised machine learning techniques, such as clustering, may require information not readily (or ever) available to the analyst for proper classification of most activities, since they are user-specific. For example, a patient treating a disease may visit a hospital frequently for treatment purposes, whereas a nurse may commute daily to the hospital for work. Accurately identifying these differences is important for proper design of the transport network, but it is a challenge for unsupervised learning. Assuming that no other information is available, clustering can be performed based on spatial and/or time coordinates, grouping together activities that are located close to each other and/or realized at approximately the same times. One of such algorithms is DBSCAN (Ester *et al.*, 1996), which uses the GPS coordinates along with thresholds for maximum distance (or the radius ϵ of the circle formed with a point in its center) and minimum number of points ('MinPts') to cluster points together. Similar applications have pointed out good suitability of the method for this purpose (Liu *et al.*, 2019; Bhadane and Shah, 2020; Xiong *et al.*, 2020; Marra, 2021).

DBSCAN is based on the concept of density-connectivity, which efficiently classifies points in clusters of arbitrary shape, without the need to specify an initial number of clusters. The density of an arbitrary point is defined as the number of points within a circle of radius ϵ from that point. Then, for each point in the formed clusters, the circle of radius ϵ contains the minimum number of points specified. Two main definitions are important to form the density-based notion of a cluster (Ester *et al.*, 1996): density-reachable and density-connected points. A point p is density-reachable from a point q, with respect to ϵ and 'MinPts', if there is a chain of points $p_1, \cdots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is in the ϵ neighborhood of p_i and the number of points in that neighborhood is greater of equal than 'MinPts' (in this case p_{i+1} is also called directly density-reachable from p_i and p_i is a core point). If there is a point o such that both p and q are density-reachable from o with respect to ϵ and 'MinPts', then p and q are said to be density-connected with respect to ϵ and 'MinPts'. Density-connectivity is a symmetric relation, so that two points (called borders points) can belong to the same cluster even without sharing a common core point, but then it must be the case that there exists a common core point from which these borders points are density-reachable.

Although DBSCAN automatically defines the number of clusters based on the specified inputs ϵ and 'MinPts', one of its biggest drawbacks lies in the fact that all the clusters are based solely on these two parameters, so if the data has points forming clusters of varying densities, the resulting clusters may be meaningless. In this case, a trade-off between accuracy and detail of activities is necessary for the application on GPS Tracking data. This trade-off entails the unknown nature of the activities, but assumes some common behavioural patterns, such as the ones involving home and work locations. A simple rule of thumb is to define the cluster representing 'home' as the one with the most points (i.e. location with the highest number of activities), the cluster representing 'work' as the second most visited one, and all other clusters just being assigned the label 'other'. Temporal information can also be included, as in Marra (2021), where 'home' and 'work' (or 'school' in case of students) classifications were constrained to the clusters with the highest number of activities during nighttime and daytime, respectively, and only on weekdays. With respect to the other activities, clearly, even in the absence of labelled data, better classification could be pursued. For example, matching of GPS coordinates with locations of commercial buildings, such as gyms and groceries stores, could shed light on some behavioural characteristics of the users. The aforementioned trade-off comes into evidence in this problem, as aiming for a better classification (more labels) of activities comes at the cost of losing accuracy (i.e. the algorithm detects more activities, but also makes more mistakes). Clearly, even the rule-of-thumb-based approach is subject to inaccuracies, for example if the user's mobility behaviour does not correspond to the assumed one (e.g. a high frequency of 'home office' activity may cause the actual work location not to be properly identified as such). Hence, if more labels of activities are needed, the analyst has to deal with the inaccuracies that inevitably appear. For the purpose of this paper, the three-level classification was considered enough to capture routine behaviours of travellers in terms of recurrent trips between home and work locations. Before studying such behavioural aspects, a comparison between clusters formed by different parametrizations is discussed. More specifically, the choice of the distance parameter ϵ and the consideration of time variables such as 'day of the week' and 'activity time'.

The distance parameter ϵ plays an important role in the DBSCAN algorithm since it defines a maximum distance threshold for activities to be considered in the same location (hence, classified as the same activity). Intuitively, this threshold should be high enough to accommodate possible GPS location errors as well as locations big in size, but not too high so that it does not include neighbouring locations corresponding to other activities, which could cause, for instance, a supermarket in the neighbourhood to be labelled as 'home'. However, this naive type of reasoning may cause important data loss. Considering again the supermarket example, it may be the case that, occasionally, instead of going from work to home directly, the individual goes first to the neighbourhood supermarket and, then, home. Depending on the parametrized clustering distance, this trip may not be classified as 'work-home' (and, therefore, excluded for further analysis), although the same patterns of times and routes defining common 'work-home' trips are likely observed. Therefore, the distance parameters should not be arbitrarily defined, instead chosen based on the purpose of the study and, possibly, data-specific. To illustrate this choice based on the distance parameter only, and without any consideration of time variables, Fig. 1 depicts two scenarios for the activities of one of the tracked users: on the left panel, DBSCAN with ϵ (eps) set to 500m and, on the right panel, ϵ (eps) was set to 100m. The red dots correspond to the tracked activities and the activities which do not belong to either the 'home' nor the 'work' clusters are labelled with a number according to the cluster assigned by DBSCAN (notice that some activities belong to the same cluster, so some labels appear more than once). The activities belonging to the 'work' cluster are shown inside of the red circle, whereas activities belonging to the 'home' cluster are shown within the blue circle.

Figure 1: Clusters for the same user based on DBSCAN with two values of the eps parameter: 500m (left) and 100m (right).

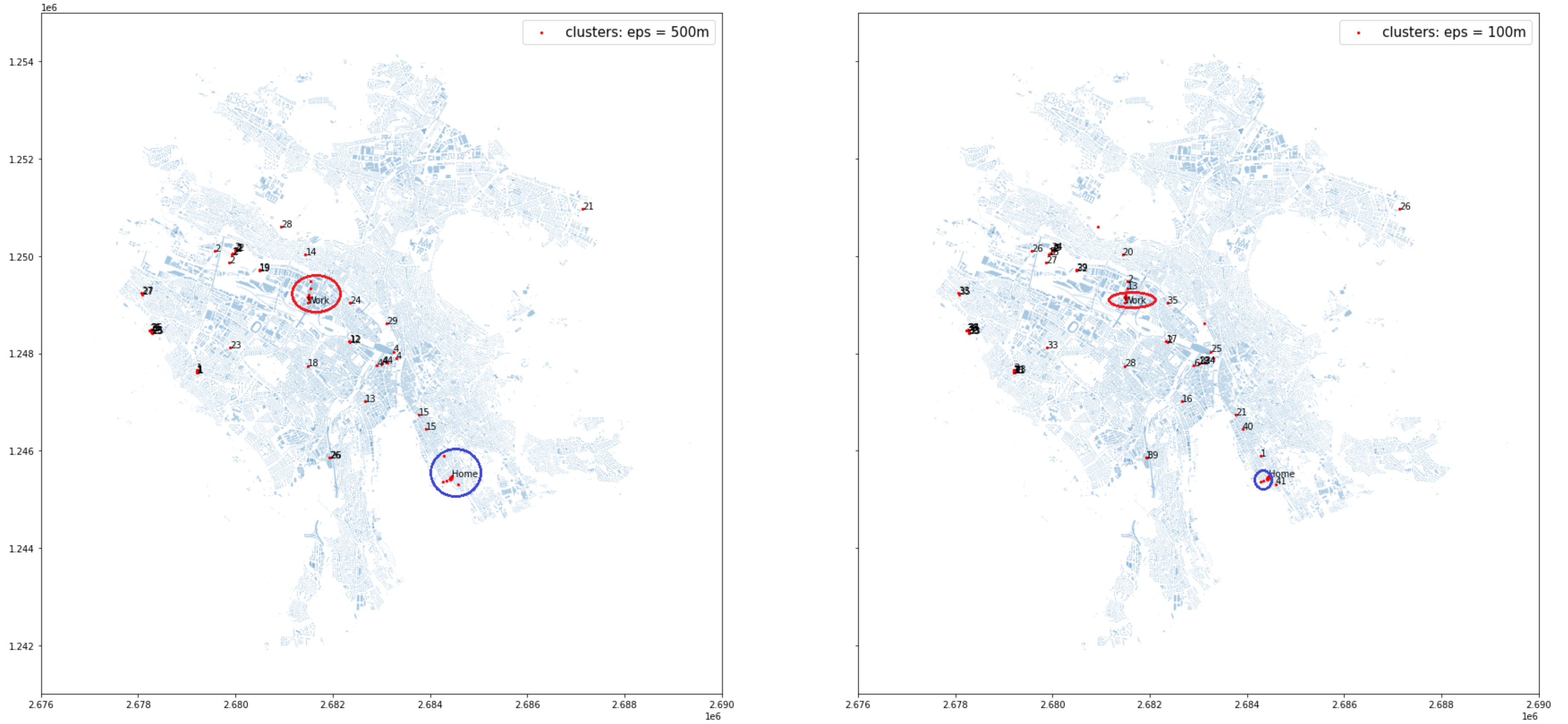


Fig. 1 puts into evidence the problem surrounding the choice of the distance parameter for unsupervised learning applications. While the clustering with $\epsilon = 100$m (right panel) seems to be doing fairly good in determining the densest clusters for ‘home’ and ‘work’ locations, some activities nearby these locations are left behind and classified in single clusters (e.g. cluster 41 in the surroundings of ‘home’ and clusters 2 and 13 in the surroundings of ‘work’). Although these seem to be, in fact, different locations, they could still serve as data for the purpose of identifying route patterns of the individual commuting through these locations, as they are still close in distance. Another interesting observation comes from the activities in clusters 2 and 4 when DBSCAN is applied with $\epsilon = 500$m (left panel). While these two clusters are well-defined and separated from others in the left panel, these same activities are classified in many different (isolated) clusters in the right panel. Although the different activities in cluster 4 seem to correspond to different locations surrounding Zurich’s main station, the different activities in cluster 2 correspond to the same location: a city park. This means that setting $\epsilon = 100$m potentially misclassifies ‘home’ and ‘work’ activities if these places are big in dimension.

Introducing two time variables, namely the day of the week and the activity time, as suggested in Marra (2021), leads to a different clustering configuration. After implementing DBSCAN for a given ϵ, the time variables are considered. ‘Home’ is the cluster with the highest number of activities (weighted by their duration) during weekdays, between 23:00 and 06:00, whereas ‘work’ is the cluster with the highest number of activities (weighted by their duration) during weekdays, between 09:00-12:00 and 13:00-17:00. Table 1 shows a comparison, for different values of the parameter ϵ, of the number of activities, from a total of 15265, classified in the home and work clusters for the strategy considering only the (spatial) distance variable and also for the strategy described in Marra (2021), that utilises both distance and time parameters.

Table 1: Comparison between clustering based on distance parameters only and clustering based on distance and time parameters for different values of DBSCAN ϵ parameter.

ϵ [m]	50	100	150	200	250	300	350	400	450	500	1000	1500
Number of Work activities - distance only	2300	2524	2652	2707	2781	2825	2935	2976	3039	3111	2986	2816
Number of Home activities - distance only	4526	4827	4948	5031	5120	5226	5348	5482	5584	5738	7586	8797
Number of Work activities - distance and time	1966	2121	2273	2314	2375	2391	2486	2514	2552	2611	2490	2111
Number of Home activities - distance and time	4189	4526	4658	4743	4822	4923	5045	5162	5269	5382	7021	8291
Number of IsMatch Work	1506	1672	1728	1670	1708	1739	1788	1779	1804	1808	1408	1234
% IsMatch Work	65%	66%	65%	62%	61%	62%	61%	60%	59%	58%	47%	44%
Number of IsMatch Home	3931	4281	4403	4433	4531	4648	4707	4815	4877	4930	6437	7815
% IsMatch Home	87%	89%	89%	88%	88%	89%	88%	88%	87%	86%	85%	89%

Table 1 reveals that, as expected, the number of activities classified in each cluster (home or work), in general, slightly increases as the distance parameter ϵ increases from 50m until about 500m. More interestingly, however, is to observe that in the clustering obtained by using only the distance parameter (first two rows), for which the most visited location was considered as home and the second most visited location was considered as work, there is a tipping point somewhere between $\epsilon = 500$m and $\epsilon = 1000$m, where activities classified as work move to the home cluster, i.e. the home cluster starts to absorb activities that were before classified as work, or other. In fact, for a total of 15265 activities, when ϵ is set to 1000m, roughly 50% of the activities are classified as home, and this percentage increases to 58% when $\epsilon = 1500$m is considered (*versus* 20% and 18%, respectively, for work activities). Moreover, the number of activities classified as home increases 27% from $\epsilon = 50$m to $\epsilon = 500$m, and work activities, for the same ϵ interval, increase about 35%. However, from $\epsilon = 500$m to $\epsilon = 1500$m, home activities increase 53% (a total increase of 94% from $\epsilon = 50$m to $\epsilon = 1500$m), whereas work activities decrease by 9% (although still keeping a positive total increase of 22% in the whole interval from $\epsilon = 50$m to $\epsilon =$ 1500m). This suggests some behavioural patterns of the individuals, the most prominent one being that individuals tend to choose locations for their activities that are close to where they live. The same does not seem to be true for work locations, for which the number of activities increase slightly every 50m, indicating that there are a few activities locations close to work that the user may visit, but when ϵ is big enough (indicating locations that are not within walkable distance) the cluster starts to lose activities for the home cluster, or it can even become part of the home cluster, in which case another activity will be mistakenly classified as work. Of course, this type of analysis involving changes in categories does not offer the full details of what happens at an individual level. As mentioned, it could happen, for instance, that the work location is very close to the home location, and increasing the distance parameter by a few meters would cause home and work to be classified together as home, which, in turn, causes another activity to be mistakenly classified as work.

The results of the inclusion of the two time variables (time of the activity and day of the week) in the clustering analysis, as suggested by Marra (2021), are presented in rows three and four in Table 1. For this analysis, it is important to mention that the number of activities considered is still the same, and the clusters still come from DBSCAN applied with the same parameters. The difference here is that home and work are not taken to be the first and second, respectively, most occurring activities. Instead, a subset of the obtained clusters, filtered by weekdays, is considered. The cluster with the highest number of activities that were realized between 23:00 and 06:00 in this subset is classified as home, and the cluster with the highest number of activities, excluding home activities, that were realized between 09:00-12:00 and 13:00-17:00 is classified as work. Notice that an activity that was assigned a cluster number by DBSCAN, say 1, regardless of its realization time, may still be considered as home if the other activities that were also assigned the number 1 by DBSCAN still form the majority of activities realized between 23:00 and 06:00.

For this type of clustering strategy, the numbers are different from the previous strategy, although the pattern of the changes (as measured by the sign and amplitudes of the percentages) behaves similarly. In particular, the number of home activities increases 28% from $\epsilon = 50$m to $\epsilon = 500$m, and then 54% from $\epsilon = 500$m to $\epsilon = 1500$m (a total increase of 98% in the interval from $\epsilon = 50$m to $\epsilon = 1500$m). The percentages for work are 33%, -19%, 7%, meaning that the two extremes ϵ for work activities lead approximately to the same number of activities. At $\epsilon = 1000$m, about 46% of the activities are classified as home, and this percentage increases to 54% when $\epsilon = 1500$m is considered (*versus* 16% and 14%, respectively, for work activities), which is close to what was obtained with the distance-only strategy. However, more interesting than comparing the percentages is to compare the percentages for activities whose classification matches in both strategies. The percentages shown in rows '% IsMatch Work' and '% IsMatch Home' represent the percentages of agreement between the activities that are classified in the same cluster in both strategies over the number of activities in the cluster according to the first strategy (distance only). The agreement for work activities starts at around 65% and slowly decreases to 44% as ϵ increases, and, for home, the percentages of agreement are stable (and high) throughout the whole interval considered, ranging from about 85% to 89% (it reaches its maximum when ϵ is set to 1500m). Hence, the percentages for home reveal a good agreement, and it improves as the distance parameter gets bigger or, in other words, when the home cluster is broader. The same does not occur for the cluster of work activities, for which the agreement is higher under lower distances or under more compact clusters. In other words, constraining the activities to a certain time interval and to be at weekdays affects the classification of work activities more than it does to home activities, when compared to the clustering obtained solely with the distance parameter.

Putting this into another perspective, it means that home, which is taken as the most visited location, is more robust against variations of distance and time parameters than work from a clustering perspective, i.e., it is easier to classify a set of activities that includes the actual home location than it is to classify work correctly. Aside from the assumption that an individual will visit home the most, this also reveals a bias for individuals towards choosing locations for activities that are nearby their homes, forming a dense cluster around home location. On the other hand, correct classification of work activities may be tied to the correctness of the chosen distance and time interval for the population under study, although this relationship is not clear and can only be assumed in the absence of labelled data. For example, in this study, the first strategy assumed work as the second most visited location, and then the second strategy assumed work as the most visited location during weekdays and at a specific daytime window. However, as shown, the agreement of these two clustering strategies was low, reflecting a lack of an obvious spatiotemporal clustering strategy for identifying work locations.

Upon results of both strategies, the crucial question still remains: which strategy (distance-only or distance and time) is best suitable for the purpose of unsupervised clustering of work and home activities? In fact, two more experiments are conducted, both involving changes in the time variables constraints. In the first one, the interval for home activities is modified from 19:00 to 07:00 instead of 23:00 to 06:00. In the second one, the original intervals are kept for both work and home, but now all days of the week (and not only weekdays) are considered. No changes are made to the time intervals for work activities, first to avoid possible intersections with the time interval for home and, second, because other time intervals do not make practical sense considering the population under study. Fig. 2 depicts the results of all strategies in terms of the percentages of agreements when compared to the distance-only strategy.

Figure 2: % of agreement with distance-only strategy

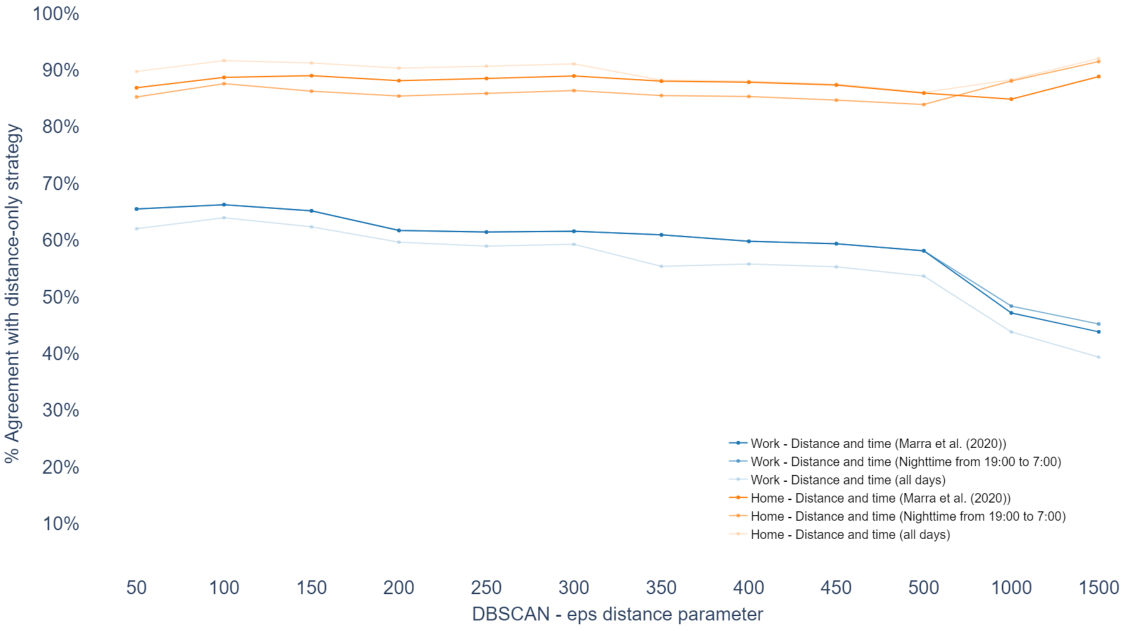


Fig. 2 shows that different strategies involving inclusion of time variables lead to similar results in terms of percentages of agreement with the original (distance-only) strategy. The agreement is always higher for home activities (mean of 87.9%, standard deviation of 2.15%) with a slight increase towards the end when ϵ is greater than 1000m. For work activities, the agreement is poor (mean of 58.1%, standard deviation of 6.85%), and the percentages decrease with the increase of ϵ, a effect that is more pronounced when ϵ is greater than 500m. In particular, when comparing across strategies, the lowest rates of agreement are obtained for work activities when relaxing the constraint on time to all days (the light blue line below all other curves), instead of only weekdays, even though the interval between 09:00-12:00 and 13:00-17:00 remains the same. In practice, this analysis reveals that, especially when the clustering radius ϵ is high, adding a time constraint causes the cluster not to correspond to the second most visited location (according to the distance-only strategy) nearly about half of the time. Moreover, the agreement with the second most visited location is higher when constraining to weekdays in a particular time interval, *versus* all days in the same time interval. Hence, when time variables are considered, the new clusters many times do not reflect the second most visited location, especially if all days are considered. From another perspective, if the clustering strategy using distance, a time interval and weekdays only is taken as the most accurate for classifying home and work, then activities labelled as work will correspond to nearly 65% of the second most visited location when $\epsilon = 50$m and 45% when $\epsilon = 1500$m. If the restriction on weekdays is removed, but the time interval between 09:00-12:00 and 13:00-17:00 is kept, then the percentages drop slightly to 62% and 39%, respectively, showing that the variable time interval plays a bigger role than the variable days of the week in the classification strategy. Therefore, for a lot of travellers, there seems to be a mismatch between the second most visited location and the location that is the most visited during weekdays at the time interval 09:00-12:00 and 13:00-17:00, which is labelled as work in the second strategy. One example could be the gym activity: an active person could easily go to the gym 6-7 times a week, and go to work 5 times a week (and say that the work schedule is flexible). When restricting the work to be at a given time interval from 09:00-12:00 and 13:00-17:00, even if not all five days are within this interval, the algorithm correctly captures the work activity. However, if all days are considered (whether in the time interval or not), then going to the gym could be a more frequent activity than work. To further compare the scenarios, some statistics are presented in Table 2, as follows:

Table 2: Comparison between clustering based on distance parameters only and clustering based on distance and time parameters for different strategies.

	Minimum		Mean		Maximum		Std. Dev.		C.V.	
Clustering Strategy	Work	Home	Work	Home	Work	Home	Work	Home	Work	Home
Distance-only	2300	4526	2804	5684	3111	8797	233	1245	0.08	0.22
Distance and time (Marra et al. (2020))	1966	4189	2350	5336	2611	8291	200	1164	0.09	0.22
Distance and time (Nighttime from 19:00 to 7:00)	1926	4136	2339	5314	2647	8488	229	1254	0.1	0.24
Distance and time (all days)	1808	4301	2144	5431	2434	8567	198	1233	0.09	0.23

Table 2 shows that, on average, using the distance-only strategy leads to a higher number of activities classified as work and home for different values of ϵ. By adding two time variables, and restricting the analysis to weekdays in a particular time interval for home and another interval for work (rows 2 and 3), the average number of activities classified in these two clusters decreases, and the numbers do not differ much if the time interval considered for home increases from 23:00-6:00 to 19:00-7:00, although the variability is slightly superior for the second interval. If the restriction on weekdays is relaxed, and only the original time intervals are considered (row 4), the average number of work activities and their variability decrease in comparison to the strategies with both variables, but the same is not observed for home, which ends up absorbing slightly more activities. Again, the observed statistics corroborate with the conclusion that the time interval has a greater effect than day of the week in determining the location of the work cluster. Moreover, by comparing the strategies including time variables (rows 2-4) with the strategy including only the distance variable (row 1), the coefficient of variation (CV), which is a statistical measure of the relative dispersion around the mean, are all very close. The results for this analysis show that inclusion of the two time variables, although not likely to have a big effect on the correct assignment of home (or surroundings) locations, has a significant impact on the work cluster.

In the complete absence of labels, the choice of a proper clustering strategy needs to be linked to behavioural assumptions or relied upon information gathered through surveys, e.g. socioeconomic questionnaires that prompt for the usual work schedule. Hence, for example, if it is assumed that workers do not go to the workplace during weekends, then adding the constraint on weekdays will lead to better accuracy for work activities classification. On an individual level, if, for instance, the user was prompted to enter a usual work schedule, then this information could serve as input for the classification algorithm. Hence, restrictions on time variables could improve the accuracy of the classification, although they are still linked to behavioural assumptions or, in the best-case scenario, attached to known information about travellers' daily routines, which must then be available.

For the *ETH-IVT Travel Diary* dataset, which did not include labels for activities, sociodemographic and behavioural data were acquired prior to the study's start date. For the 172 users, 170 questionnaires were answered. Among the questions, some are of special interest for the purpose of defining the clustering parameters: main occupation (student, worker, both or other), average time spent from home to work location (or school in case the main occupation is a student), days of the week in which the user dislocates to work (or school), ownership of PT subscription, ownership of driver's license, frequency of use of PT and private modes. These selected statistics, shown in Table 3, are used to support the choice of a proper clustering strategy.

Table 3: Sociodemographic and behavioural characteristics of travellers in the *ETH-IVT Travel Diary* survey

Variable	Category	Counts
Main Occupation	Employed / self-employed	107
	Student	41
	Student and Employed / self-employed	11
	Other	11
Average time spent from home to work [min]	$[0, 15]$	51
	$(15, 30]$	72
	$(30, 45]$	25
	> 45	21
Work/School days	Weekdays Only	145
	Everyday	10
	Weekdays and Saturday	7
Driver's license ownership	Yes	119
	No	50
PT subscription	Yes	127
	No	38
Frequency of use (PT and Private Modes)	PT = almost daily; bike/car/private = 1-3 days per week	58
	PT = almost daily; bike/car/private = 1-3 days per month	26
	PT = almost daily; bike/car/private = almost daily	20
	PT = 1-3 days per week; bike/car/private = almost daily	14
	PT = almost daily; bike/car/private = rarely to never	13
	PT = 1-3 days per week; bike/car/private = 1-3 days per week	12
	PT = 1-3 days per month; bike/car/private = almost daily	8
	PT = 1-3 days per month; bike/car/private = 1-3 days per month	5
	PT = 1-3 days per week; bike/car/private = 1-3 days per month	4
	PT = 1-3 days per month; bike/car/private = rarely to never	3
	PT = 1-3 days per month; bike/car/private = 1-3 days per week	2
	PT = 1-3 days per week; bike/car/private = rarely to never	2
	PT = rarely to never; bike/car/private = almost daily	1
	PT = rarely to never; bike/car/private = 1-3 days per week	1
	PT = rarely to never; bike/car/private = rarely to never	1

Table 3 shows that the majority of the survey's participants is an employee or self-employed (about 63%), although a significant part is a student (24%) or both worker and student (6.5%), who dislocates to work (or school) on weekdays only (85%) and for which home-work trips take, on average, between 0 and 30 minutes (72%). Moreover, about 75% of the respondents declared owning a public transport subscription, whereas 70% own a driver's license. In terms of transport modes, 69% declared to use PT (bus, train or tram) on a daily basis and 30% about 1-3 days per week. Private modes (taxi, uber, rental bikes), or own car/bike are used on a daily basis by about 25% of the respondents, 1-3 days per week by 43% and 1-3 days per month by 21%. Hence, for the individuals in the survey, it can be inferred that the majority of them lives relatively close to the work/school place, owns a PT subscription, uses PT as the main dislocation mode for work/school activities and commutes to work/school on a weekday basis. These aspects support the use of a small distance parameter for clustering, since work and home may be close as suggested by the average dislocation times, and the use of the two time variables, namely day of the week restricted to 'weekdays' and time interval for work/school activities ranging from 09:00-12:00 and 13:00-17:00. Although the survey did not acquire details about studying or working times, the time interval can be chosen to reflect the most common in the area where the survey was taken, in this case the city of Zürich. As per home time interval, Switzerland's law prohibits working on Sundays and at night from 23:00 to 6:00. Only some sectors are not subject to this ban, meaning that this is an appropriate interval to be designated for home activities, as the majority of services will be closed. In practical terms, this means that the parametrization suggested in Marra (2021) is the one considered the most appropriate to distinguish between home and work activities and, hence, to identify commuting trips from home to work and vice-versa. The next section investigates these trips.

4 Behavioural Aspects of Commuting Trips

The *ETH-IVT Travel Diary* survey indicates that most travellers usually commute on a weekday basis from home to work (and vice-versa), and PT is used by almost the whole sample under study in a weekly basis. Therefore, understanding travellers' behaviour in these commuting trips is a key factor to enhance urban transport planning. This understanding involves contrasting travellers' choices in terms of routes with the other (PT) choices available, including possible disruptions that may have happened. Some aspects are of special interest and are further investigated in this section, including mode share, recurrence of the chosen mode of transport and route for home-work and for work-home trips, average time of the chosen route *vs.* average time of the shortest path (timetable and actual times), statistics on main time parameters (transfer times, walk times and times on each PT mode), information available to travellers in commuting trips (do travellers know about possible disruptions and do they change their route as a result?), frequency of nontrivial route choices (as defined by routes that are not in the timetable choice set or realized choice set). An important remark is that the travel diaries were restricted to PT trips.

To extract the commuting trips from unlabelled GPS data, the clustering strategy discussed in the previous section is used. First, the DBSCAN algorithm with the distance parameter ϵ set to 100m is applied to form the clusters, then the two time variables are considered, namely days of the week and time. According to this strategy, home is the cluster with the highest number of activities (weighted by their duration) during weekdays, between 23:00 and 06:00, and work is the cluster with the highest number of activities (weighted by their duration) during weekdays, between 09:00-12:00 and 13:00-17:00. The remaining clusters are labelled as 'other' and are not considered relevant for the purpose of analyzing commuting trips given the characteristics of the survey. Finally, after labelling home and work locations, commuting trips from home to work and vice-versa are obtained by identifying trips with origin and destination points in these locations. Table 4 summarizes the results obtained for activities and trips by using this clustering strategy.

Table 4: Summary of activities according to chosen clustering strategy.

	Values	%
Total Activities	15265	100.0%
Home Activities	4526	29.6%
Work Activities	2121	13.9%
Other Activities	8618	56.5%
Total Trips	2909	100.0%
Total Home-Work Trips	361	12.4%
Total Work-Home Trips	229	7.9%
Total Commuting Trips	590	20.3%
Avg. Time Trip [minutes]	22.76	-
Avg. Time Home-Work Trip [minutes]	28.15	-
Avg. Time Work-Home Trip [minutes]	30.98	-
Avg. Time Commuting Trip [minutes]	29.25	-

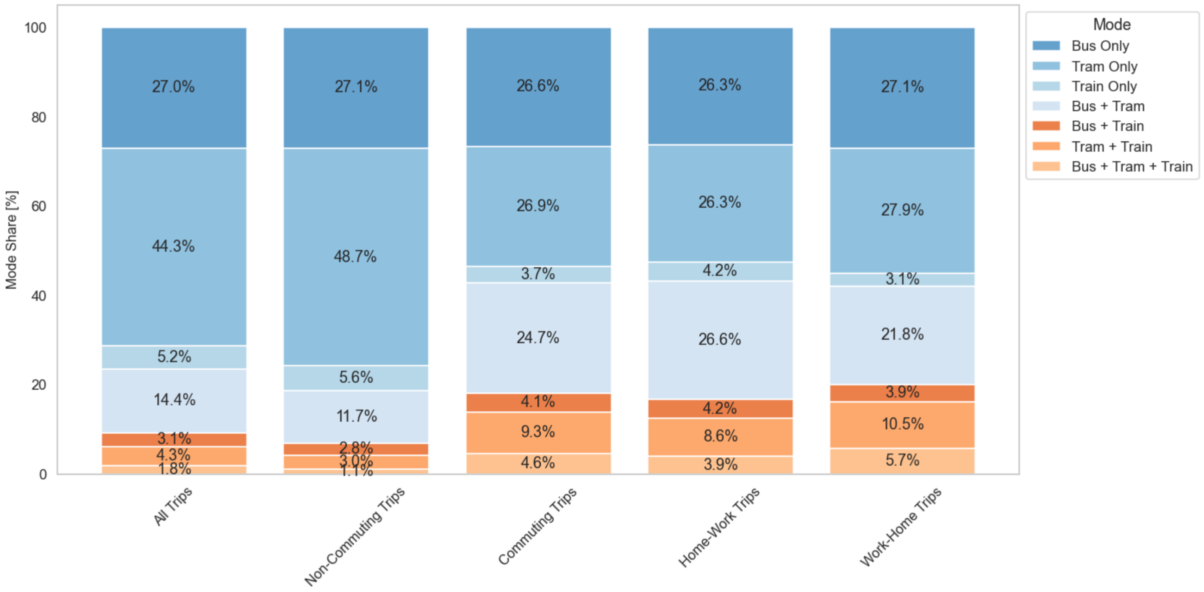
Table 4 reveals that about 20% of the trips are commuting trips from home to work, and vice-versa, with a clear imbalance between them, with home-work trips accounting for about 61% of these commuting trips. Interesting is to notice that commuting trips, on average, take longer than the average of all trips (29.25 minutes *vs.* 22.76 minutes), suggesting that many other activities are performed close to the home location. In particular, work-home trips take the longest and are, on average, slower than home-work trips (30.98 minutes *vs.* 28.15 minutes). Since the clustering distance parameter was chosen to be restrictive (100m), it is unlikely that the explanation for the difference comes from stops to other locations in the neighbourhood that are yet farther away than home. Instead, two other hypotheses seem more plausible. The first one is that traffic congestion at the end of the working hours (evening rush-hours) is more intense, making work-home trips longer than home-work trips. The second one, a seemingly intuitive explanation, is that commuters do not always aim for the shortest time trips when coming back to home from work and, for instance, may privilege some other aspects related to comfort or even longer walks, resulting in bigger times. Regarding the first hypothesis, according to traffic flow statistics in the city of Zürich (TomTom, 2021), the weekly average congestion levels are slightly worse in the evening rush hours (from 16:00 to 18:00) than in the morning rush hours (from 6:00 to 8:00), with an extra time spent driving estimated as 13 minutes (for every 30 minutes driving) in the morning rush hour *vs.* 17 minutes in the evening rush hour, a 31% increase. For evaluating the plausibility of the second hypothesis, and to investigate other aspects of commuting trips, in particular, travellers' behaviour when in commuting trips, the next subsections present several analyses made on the *ETH-IVT Travel Diary* dataset.

For the following analyzes, the algorithms and methods developed in Marra *et al.* (2019); Marra and Corman (2020a) for mode detection and choice set generation are used. In particular, the mode detection algorithm identifies activities, trips, stages and transport modes used, including information on the line, the corresponding vehicle, as well as departure and arrival stops and times. This mode detection algorithm was validated against labelled data showing an accuracy of 86.14%, so that the trips assigned to each travel diary in the dataset are assumed correct. For the CS generation algorithm, AVL data of the city of Zürich PT network is exploited to provide not only alternatives in terms of (scheduled) timetable, but also realized alternatives considering actual departure times that account for possible disruptions. In terms of the CS route options, a state of the art method for route choice in PT is used, namely the Path Size Logit Model (Marra and Corman, 2020a), which is an extension of the Multinomial Logit model that includes a penalizing factor, the Path Size, in the utility function. For further information on the model and its limitations, the reader should refer to Marra and Corman (2020a).

4.1 Mode share in commuting trips *vs* non-commuting trips

Fig. 3 shows the PT mode share for the *ETH-IVT Travel Diary* dataset, where the first bar corresponds to all trips, the second bar to all non-commuting trips and the third bar to all commuting trips. The commuting trips are also stratified in the fourth and fifth bars according to origin and destination (home-work or work-home, respectively).

Figure 3: Mode share for the *ETH-IVT Travel Diary* dataset: all trips, non-commuting trips and commuting trips (home-work and work-home)



From Fig. 3, there seems to exist differences in the mode shares of commuting and non-commuting trips. A multinomial hypothesis test can test whether the observed differences in proportions of the modes across these two (mutually exclusive) types of trips are statistically significant or not. Let Y_{ij} denote the Random Variable representing the number of subjects observed in trip type i, $i = 1$ (non-commuting), 2 (commuting), and travelling under mode j, $j = 1, \cdots, 7$, and let p_{ij} denote the proportions of such individuals. Then the model is represented by Eq. (1):

$$Y_{i1}, Y_{i2}, \cdots, Y_{i7} \sim \text{Multinomial}(n_{i.}, p_{i1}, p_{i2}, \cdots, p_{i7}), \qquad i = 1, 2 \tag{1}$$

where $n_{i.}$ is the total number of subjects in trip type i. Under the null hypothesis, the proportions of individuals opting for a mode j across all trips types is the same (homogeneous), so that the null hypothesis has 6 (i.e. $7-1$) parameters to be estimated. The alternative hypothesis is that the proportions are all different, and it has 6 (i.e. $(2-1)\times(7-1)$) parameters. Under the null hypothesis, the test statistic follows a χ^2 statistic with 6 degrees of freedom, for which the calculated value of 186.46 has an associated p-value < 0.000. Hence, there is not statistical evidence supporting the null hypothesis, and the hypothesis of homogeneity of mode share in non-commuting *vs.* commuting trips is rejected. Based on Fig. 3, the same test is conducted, now testing whether the proportions of modes in home-work and work-home trips are the same. The test statistic is 3.45 and has an associated p-value of 0.75, so that there is not statistical evidence to reject the null hypothesis of homogeneity between the two categories of commuting trips.

The tests reveal that the observed data supports similar mode share in commuting trips (home-work and work-home), but the same can not be inferred for non-commuting *vs.* commuting trips, where there is evidence in the data supporting heterogeneity of mode share. In particular, when inspecting Fig. 3 "Bus only" and "Tram only" are the most predominant modes in non-commuting trips, however, for commuting trips, the shares of mixed trips, and, in particular, "Bus + Tram" increase, which is associated with increased number of transfers and could indicate higher "willingness to transfer" aiming at faster routes in commuting trips, for example. This aspect is further investigated in the next subsection.

4.2 Recurrence of modes and routes in commuting trips

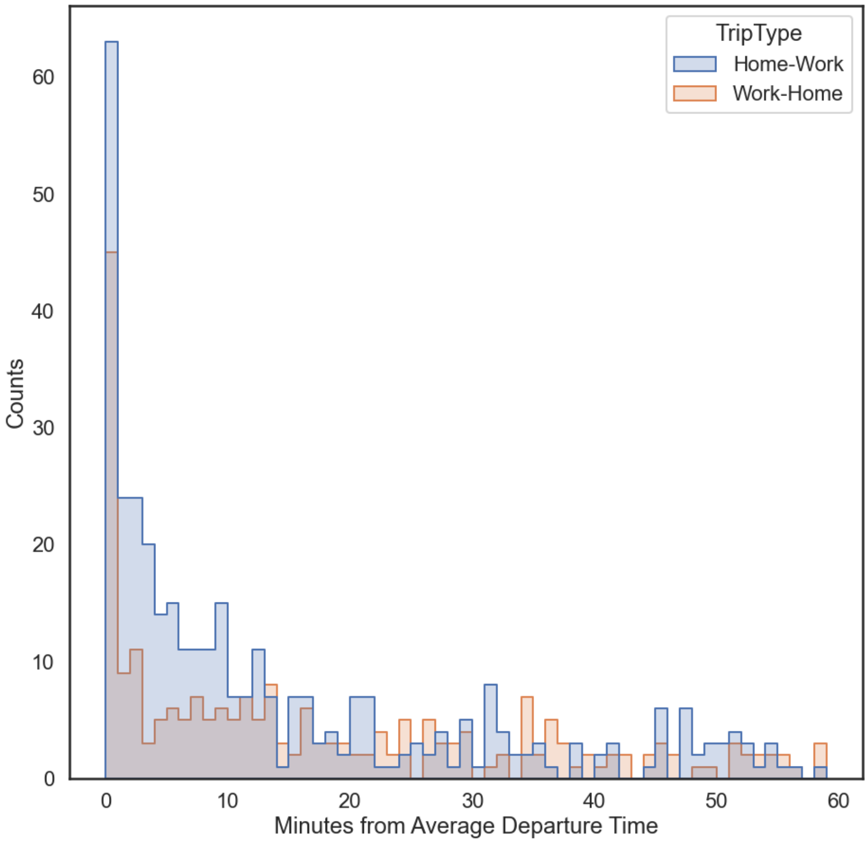
Studying the recurrence of modes and routes in PT and, particularly, the differences between choice patterns in commuting and non-commuting trips helps draw user profiles or, in other words, facilitates the understanding of what most travellers are looking for when choosing their daily routes, and which factors, if any, are somehow relaxed when the trip purpose is not directly associated with work.

The recurrence is investigated from four perspectives: departure time, mode choice, line choice and trip duration. Then, a final analysis is performed by considering all these factors at the same time.

– Recurrence in Departure Times

Fig. 4 shows the histogram of distance in minutes from average departure times, where each bin corresponds to one minute. To calculate the time distance, for each user with at least one trip in the category "Home-Work" or "Work-Home", the average departure time for all that user's trips was taken, and the absolute difference (in minutes) from each trip to the average time was calculated. To avoid possible biases due to, for example, an individual that goes back home every day for lunch and comes back to work afterwards, three shifts were considered: morning (from 6:00 to 12:00), afternoon (from 12:00 to 18:00) and other (all other times). In this case, averages were taken for each shift, and absolute differences were calculated based on these averages.

Figure 4: Histogram of distance in minutes from average departure times



For commuting trips, independently of the category, no differences from the average departure times over 60 minutes were found, meaning that departure times for home-work and work-Home trips (given a correction according to morning/afternoon/other shift) always occurred within one hour of the average time for all user's trips in the same category. In particular, over half of the trips occurred within 15 minutes of the average departure time (69%, and 58% for home-work and work-home, respectively), indicating a pattern of regularity of departure times in commuting trips, especially in the home-work category.

– Recurrence in Mode Choice

Recurrence in mode choice is first analyzed from the perspective of frequency of the most utilised (PT) mode for each individual. This differs from the mode share analysis since only the mode with the highest frequency for each individual is considered, and not the mode share of all trips. Fig. 5 shows, for all individuals, the frequencies associated with their preferred mode of transport depending on the type of trip. Tram was the preferred mode of transport for trips between home and work, and vice-versa, directly followed by bus and by the combination of bus and tram. For non-commuting trips, the percentages change dramatically, with tram-only and bus-only trips adding up to over 90%, indicating a strong preference of users for these modes of transport.

Figure 5: Mode Share of Preferred (PT) Mode of Transport for each Individual in Commuting Trips

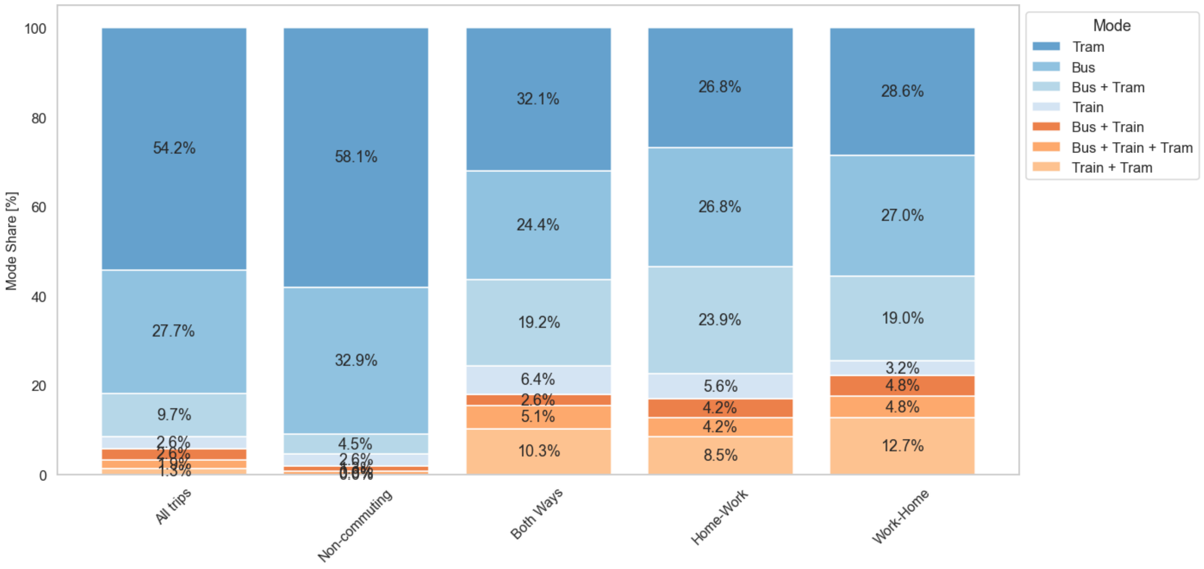
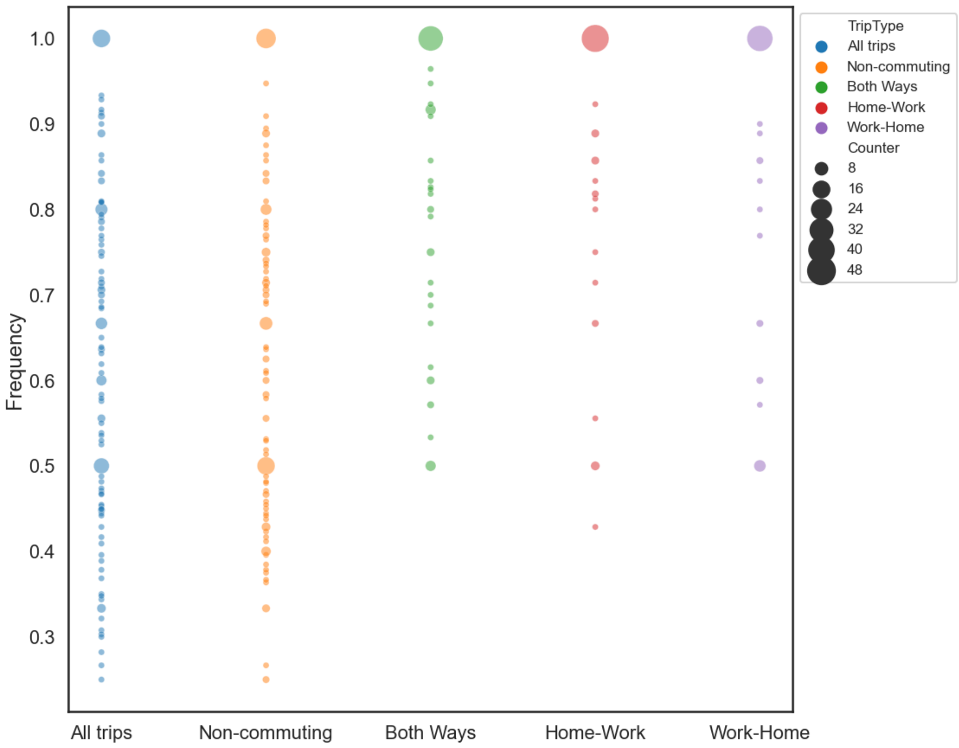


Fig. 5 reveals the preferred ("most recurrent") PT mode, but it lacks information on how recurrent it was (or the percentage/frequency in which each individual utilised that mode). In fact, the average percentage of utilisation of the preferred mode is as high as 87% for both ways trips (meaning that, on average, 87% of the observed commuting trips for each user are made with their preferred mode of transport), with a standard deviation of 17%. The averages for home-work and work-home are 92% and 89%, respectively, whereas for non-commuting trips the average is 66%, and the standard deviations are 15%, 18%, and 21%, respectively.

In Fig. 6 each circle represents the counts of users with the same frequency for their preferred (PT) mode of transport in a given trip category. The size of the circle indicates the counts, and it ranges from 1 (the smallest circles in each category) to 49 (the big red circle in the Home-Work category). For example, the big red circle at the frequency of 1 corresponds to the counts (49) of users for which the preferred mode of transport was used in all home-work trips (a frequency of 1 implies no other modes were used, so the travellers opted for the same mode of transport for all trips of the same category). Fig. 6 shows that users opting for always using their preferred mode of transport were the majority in commuting trips, as most data points fall above the 0.7 frequency, with a great concentration of users inside the big circles corresponding to frequencies of 1.0, or 100%. In particular, the percentage of users for which the preferred mode was always used was 51% for commuting trips (69% for home-work trips and 68% for work-home trips) and only 16% for non-commuting trips. Moreover, the frequency of 0.5 is of special interest, as it indicates that the preferred mode of transport is used for roughly half of the trips, so that the other half could be a mix of other modes or just one more mode, the latter being linked to users which consistently need at least one transfer (and switching of mode) to complete their journey.

Figure 6: Counts of Frequencies of the Preferred (PT) Mode of Transport per Trip Category



– Recurrence in Line Choice

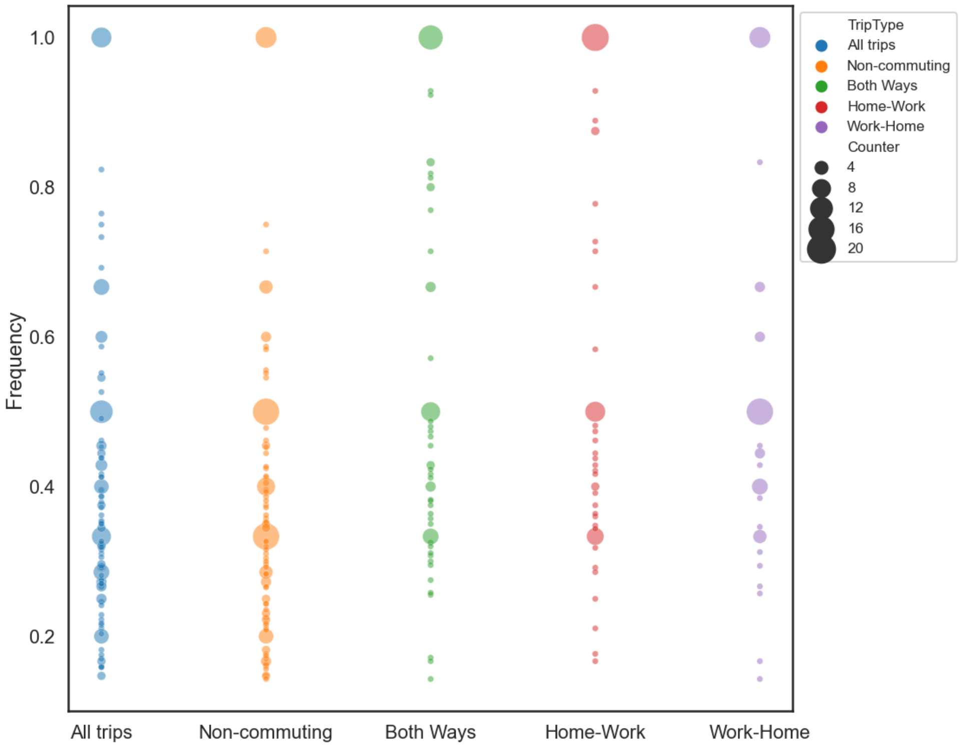
In addition to investigating the preferred mode choice, line choice preference is also investigated. There is an obvious hierarchical correlation between the two of them (the line choice is always associated with a particular mode, although one mode may have many lines), but assessing the frequency and regularity of a line choice reveals important user characteristics.

One such characteristic can be studied by contrasting the average number of lines per trip and the average unique number of lines per traveller. For non-commuting trips, each trip had an average of 1.39 lines, and each traveller used an average of 7.01 unique line numbers. For commuting trips, the average number of lines per trip was higher, 1.88, although the average unique number of lines per user was much lower, only 4.03 (home-work trips 1.89 and 3.14, work-home trips 1.85 and 3.17, for the average number of lines per trip and average unique number of lines per user, respectively). This reveals that, although travellers used, on average, more lines per trip when in commuting trips, their choice of lines was restricted to a lower number of lines. On top of that, most trips (68%) were made with only one line in non-commuting trips, whereas for commuting trips this percentage was 37%, so most trips had at least one line transfer (similar percentages were obtained for home-work and work-home trips, or 34% and 38%, respectively).

The slight discrepancy observed among the average unique number of lines per traveller in commuting trips and the subcategories home-work and work-home suggests the investigation of whether different lines are being used between these two OD pairs. In fact, for travellers having at least one home-work and one work-home trip, only for about 50% of the cases an agreement of lines between both ways was obtained. In other words, although it is possible to take the same line both ways, only about half of the travellers had at least one trip home-work and one trip work-home using the same line. For the other half of passengers in commuting trips, the lines for home-work and work-home trips never matched.

Similarly to the recurrence in mode share analysis, the frequency of the most observed line per traveller is obtained for each trip type. In Fig. 7 each circle represents the counts of users with the same frequency for their preferred (PT) line in a given trip category. The size of the circle indicates the count, and it ranges from 1 (the smallest circle in each category) to 20 (the big red circle in the Home-Work category). Hence, for instance, if the frequency is 1.0, then all trips are made with the same line (also implying the same mode). Contrasting Fig. 6 with Fig. 7 is particularly interesting. While Fig. 6 shows big circles only around frequencies of 1, Fig. 7 shows many such circles in other frequencies, especially 0.5 and 0.34, frequencies that can be linked to, at least, one or two transfers, respectively. This indicates that, although the preferred mode is used in a significant number of trips, especially in commuting ones, transfers between lines are common. For commuting trips, since most of them are based on the preferred mode, the relatively smaller circles around frequencies of 1 in Fig. 7 indicate that transfers within the same mode happen often. In fact, if, for each user, the difference between the frequency associated with the preferred mode and the frequency associated with the preferred line is considered, the average value (across all users) for this difference is 0.30 for home-work trips (standard deviation of 0.25) and 0.33 for work-home trips (standard deviation of 0.25). In practice, since the frequency associated with the preferred line will always be smaller or equal to the frequency associated with the preferred mode, this difference reveals a pattern of transferring between lines within the same mode. While a difference of zero indicates no line transfers within the mode, the higher the difference, the higher the number of transfers (within the preferred mode) that are being taken. For example, if the frequency associated with mode is 1 and the frequency associated with line is 0.67, the difference of 0.33 indicates that at least one other line (within the same mode) is being used with a frequency of 0.33, although multiple lines are also possible. Higher differences (above 0.5) imply multiple lines. Taking the example again, if the frequency associated with the preferred line was 0.44, then there would be at least two other lines filling the gap of the remaining 0.56.

Figure 7: Counts of Frequencies of the Preferred (PT) Line per Trip Category



About 20% of all commuting trips and about 19% of work-home trips are made with the use of a single line and, for these cases, no transfers are made. Home-work trips are made with only one line for about 28% of the travellers. For non-commuting trips, as expected, this percentage is much lower (about 7%), although the main reason lies in the fact that non-commuting trips incorporate trips to all OD pairs not classified as home or work, so the concentration of circles at small frequencies are not directly related to the fact that users opt for multiple lines in one journey. In general, for commuting trips, the big circles around the frequency of 1.0 indicate that many users stick to one line option when on commuting trips, although the concentration of several circles below the frequency of 0.5 indicates a significant combination of lines (transfers) in the route.

– Recurrence in Trip Duration

Trip duration, including the duration of walks, transfers and modes, is a key element to understand the regularity of commuting trips. First, regularity of duration for the same OD pair indicates that the commuters choose routes that fit their expectation (or limit) for arrival time at their destination. Second, when the use of the same line for both ways is possible, investigating differences in the duration between the two OD pairs (home-work and work-home) may reveal travellers' behavioural characteristics linked to comfort, willingness to transfer and to walk. For example, if the duration of the home-work trips is significantly shorter than work-home trips, and that is due to an extra transfer in the first category and prolonged walk in the second, then the traveller behaviour can be assumed to change from one OD pair to the other. In theory, since the locations of home and work are fixed, one would expect the trip duration for both ways to be about the same. Of course, external factors, such as traffic, crowding and weather, could have an impact on the route, e.g., traveller decides to walk instead of taking the second transfer in PT because of crowding during evening rush hour. For this study case, the proposed investigation does not consider the influence of such external factors, instead, the focus is on the regularity of the route choice. In this sense, the factors that have a recurrent impact (and are not only isolated or seldomly observed cases, like network disruptions) are assumed as known by the traveller, so that the observed daily route is a result of the traveller's behaviour for that type of trip and prior knowledge of the internal (timetable, route, etc.) and external (usual traffic and crowding conditions) factors. This should not be taken as a conservative assumption, since it is fairly general for the travellers to know the routes that best accommodate them when on commuting trips as shown by many previous works (Ma *et al.*, 2013; Goulet-Langlois *et al.*, 2016; Ortega-Tong, 2013; Zhou *et al.*, 2014; Kusakabe and Asakura, 2014).

Figure 8: Box Plots: Breakdown of Trip Times (Home-Work *vs.* Work-Home)

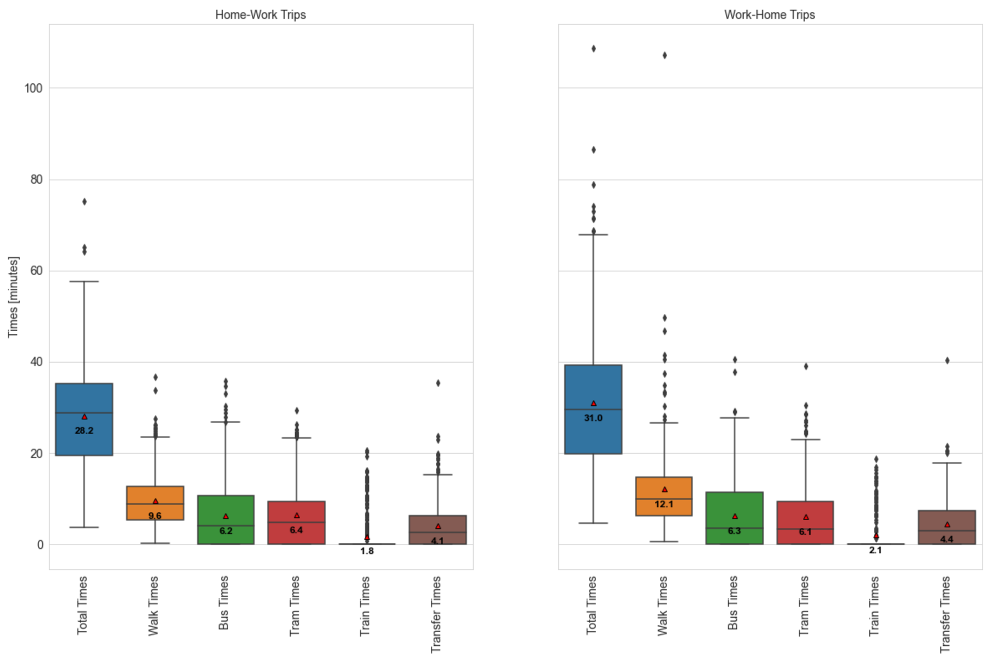


Fig. 8 shows a comparison across the commuting trips (home-work and work-home) of walk, transfer, bus, train and tram times. The times in Fig. 8 are all given in minutes. The averages are labelled for each trip category and identified by the red triangles inside the box plots. On average, work-home trips take about three more minutes to be completed in comparison to home-work trips, with longer average walking times and transfer times (12.1 and 4.4 minutes *vs.* 9.6 and 4.1 minutes, respectively). Similar average times are obtained when summing over the three PT modes (14.5 and 14.4 minutes, respectively). However, when looking into the box plots in detail, the total average time for work-home trips is clearly being influenced by some outliers in the walking times category. In fact, the median total trip times are similar (28.8 and 29.5 minutes, for home-work and work-home trips, respectively), as well as the median walking times (8.9 and 10.0 minutes, respectively) and transfer times (2.6 and 3.0, respectively), although still slightly greater for work-home trips. For the upper (third) quartile or, equivalently, the 75^{th} percentile, the differences are more pronounced, and the total time at this percentile is 35.2 minutes for home-work trips and 39.3 minutes for work-home trips, a 4.1 minutes difference, where two minutes come from differences in walking times (12.7 minutes *vs.* 14.7 minutes, respectively) and about one minute comes from transfer times (6.2 minutes *vs.* 7.4 minutes, respectively), the remaining difference resulting from PT times. The analysis reveals that up to the median (or 50% of the data points), commuting trips, whether from home-work or work-home, have similar times. However, for percentiles above the median, the differences in the times between the two categories become more evident, with work-home trips taking longer than home-work trips, even if outliers are not considered. In practical terms, outliers and trips with longer duration happened more often in the work-home path, with the differences arising from longer walking and transfer times. The outliers correspond mostly to trips for which the CS generation algorithm failed to find a correspondence among the first 100 options in terms of the utility function considered. The reasons behind this will be commented on subsection 4.5.

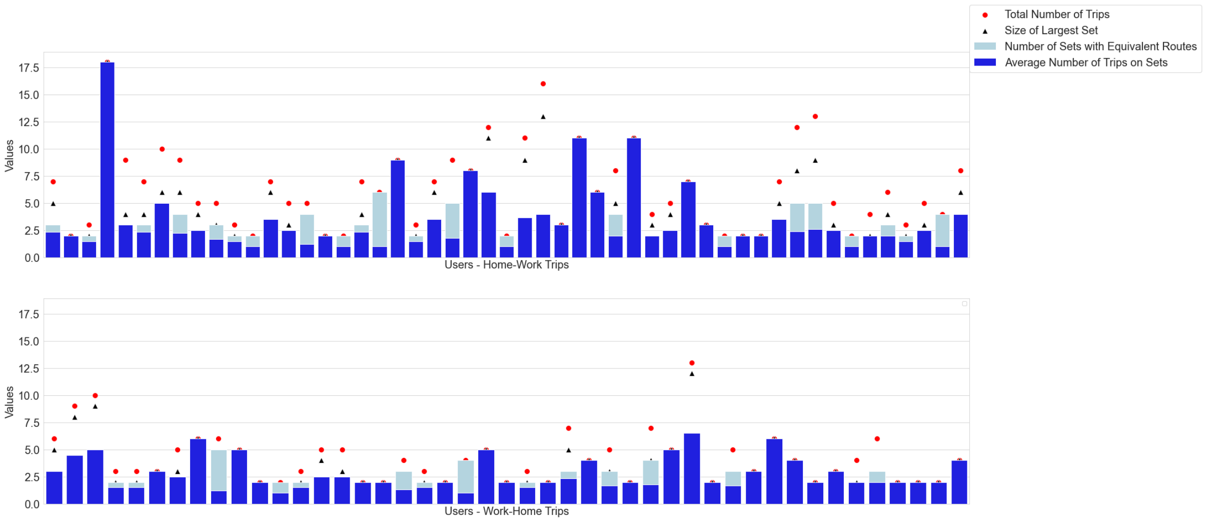
– Recurrence of all factors together

The previous subsections considered recurrence from the perspective of isolated factors, namely departure times, mode choice, line choice, and trip duration. These types of analyses allow to understand not only the relative importance and main statistics of each factor in each type of trip, but also to make comparisons, when applicable, with non-commuting trip types. The consideration of all factors together may not make practical sense for non-commuting trips, since departure times, trip duration and line choice vary significantly. However, for commuting trips, considering all these factors together reveals the proportion of these trips that can be assumed equivalent, and/or which level of similarity is realized. Hence, the interest of this subsection shifts to knowing whether, among all commuting trips for a given user, the observed choices were recurrent.

For evaluating the similarity of the trips, a similarity metric is proposed based on departure times, trip duration and line choice (mode choice is omitted to avoid redundancy since line implies mode, but not the contrary). The heuristic first groups, for all the commuting trips of a given user, the trips that are made with the same lines, and trips without a match are discarded. Then, for departure times and trip duration, trips in each subset are grouped together if they are within 5 minutes (from departure time or duration) from each other or accessible from at least one common trip. For example, a trip with a departure time at 8:00 am and another trip with a departure time at 8:10 am are initially not grouped together, but if there exists a trip with a departure time equal to 8:05 am, for example, then the 8:10 am trip is accessible from the 8:00 am trip through the 8:05 am trip, so the three of them are grouped together. The “accessibility” time rule is applied to both trip departure and trip duration variables. The output is a subset of trips, for each user, which can be considered equal from the perspective of lines, departure times and duration.

Fig. 9 shows, for each user in home-work trips (top) and work-home trips (bottom), the number of sets of trips that were grouped together according to the similarity metric (light blue bar), the average number of trips on all sets (blue bar), the total number of trips (red circle), and the size of the largest set (maximum number of trips grouped together, black triangle).

Figure 9: Similarity Metrics for Users with Home-Work and Work-Home Trips



From Fig. 9, the proximity of the red circle and the black triangle indicates the proportion of similarity in commuting trips. In other words, the closer the circle and the triangle, the higher the proportion of commuting trips corresponding to the same route, with very close departure times and duration, and made with the same lines. The proximity of the red circle and the black triangle can be used as a metric for evaluating the recurrence of such trips, by taking the percentage of trips in the largest set relative to the total number of trips for each user. Table 5 summarizes the results for users in home-work and work-home trips.

Table 5: Statistics for the proportion of similar trips across all users in commuting trips.

	Home-Work Trips	Work-Home Trips
Users Count	51	45
Avg. Number of Trips per User	6.3	4.2
Mean	71.8%	81.3%
Std. Dev.	21.7%	23.3%
Minimum	16.7%	25.0%
25^{th} percentile	58.6%	66.7%
Median	66.7%	100%
75^{th} percentile	95.8%	100%
Maximum	100%	100%

Out of the 71 travellers with recorded home-work trips, 51 had at least two trips grouped together according to the similarity metric. Each bar in Fig. 9 corresponds to one of such travellers, with an average of 6.3 trips per traveller. Of the 63 travellers with work-home trips, 45 had at least two trips, with an average of 4.2 trips per traveller. Table 5 shows that many travellers opt for the same routes and close schedules when choosing their commuting trips. In particular, both mean and median values for the similarity metric, i.e. the ratio of the largest set of similar trips over the total number of trips, were high (means of 71.8% and 66.7%, medians of 66.7% and 100% for home-work and work-home trips, respectively). Furthermore, the blue bars (average number of trips on similarity sets) are predominant over the light blue bars (number of sets), indicating that most users stick to a preferred route and recurrent travel choices when commuting between home and work. In the case of travellers on work-home trips, roughly half of them always used the same lines and same schedules (in Fig. 9 they can be easily identified by checking the bar for which the position of the red circle coincides with the upper extremity of the blue bar). Although median values for travellers in home-work trips were not as high as 100%, many users that always used the same lines and schedules had big sets of trips, some of them exceeding 10 trips.

From the analyses made, when factors are considered isolated, important differences among aspects of home-work and work-home trips can be found, especially average departure times, different line choices depending on direction, and trip duration (with prominent differences coming mainly from transfer and walking times). However, when considering a smaller subset of those trips based on the similarity metric, the analysis revealed that there is a significant part of commuters that are consistent with their route and time choices for both home-work and work-home trips, and although their characteristics may differ (e.g. different line choices and departure times), most trips in each category follow a default choice.

The remaining of the paper is dedicated to measuring how good were the observed recurrent choices, evaluating the impact of disruptions in commuting trips and investigating route choices that have unusual walking times, duration, and number of transfers.

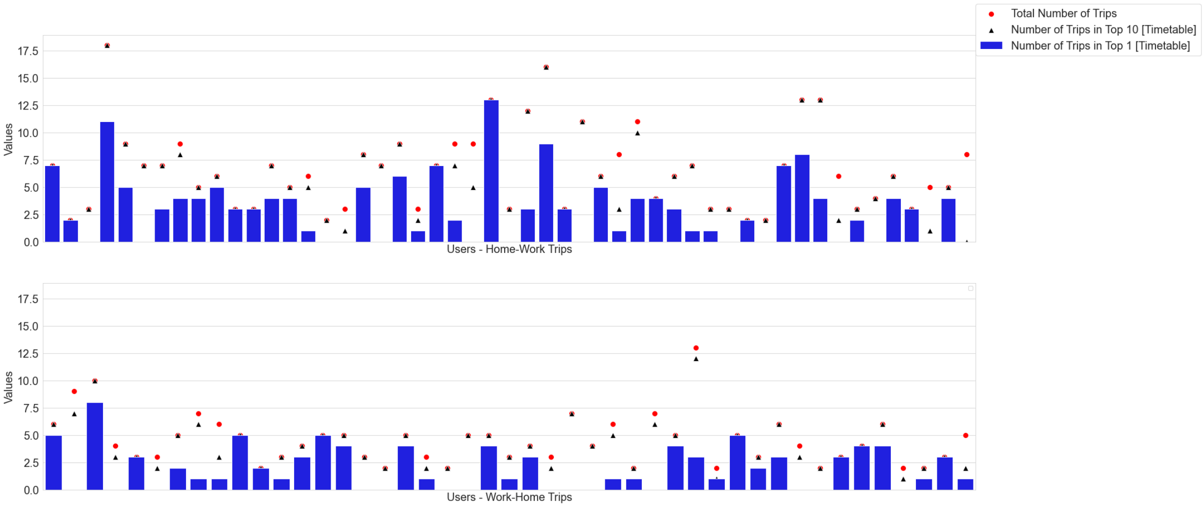
4.3 Comparison of commuting trip choices with timetable and realized times

As shown before, a significant part of the commuting trips for each traveller can be grouped together following a similarity criterion that considers line choice, departure times and total duration. For these trips, travellers choose the path that best suits them. A follow-up research question is whether these choices are actually good in terms of a common criterion. A simple, and yet frequently used one, is the shortest time path (or shortest trip duration). However, this criterion may not be realistic in the context of multimodal, strongly connected networks with short headways, as in the city of Zürich. In such cases, the fastest paths may involve a combination of multiple lines and modes, with many transfers, which is often unrealistic in practice. Hence, a utility function based on walking times, transfer times, times on PT (bus, train, tram), a penalty for the number of transfers and also a correction for overlapping paths is considered as described in Marra and Corman (2020a). Then, a simple cost function based on the travel times with a penalty of 5 minutes for each transfer is used to sort the final CS.

Using that criterion, and the CS generation algorithm to identify the available alternatives to users according to the timetable and also according to realized times, the two CS generated (for timetable and realized times) were sorted, where the first position in the CS corresponds to the route having minimum duration (considering the 5 minutes penalty for each transfer in the total duration). The path choice made by the user is then compared to the ordered positions in the CS and, for instance, if the path is in the first position, then the traveller's choice is considered optimal in terms of the criterion adopted. Moreover, a comparison between the timetable CS and the actual (realized) CS reveals important aspects regarding service reliability and the information available to travellers. For instance, in case of disruptions, a top 1 position in the timetable CS but a suboptimal position in the actual CS indicates that disruptions have happened, and the traveller stuck to the original plan, although better route choices were available.

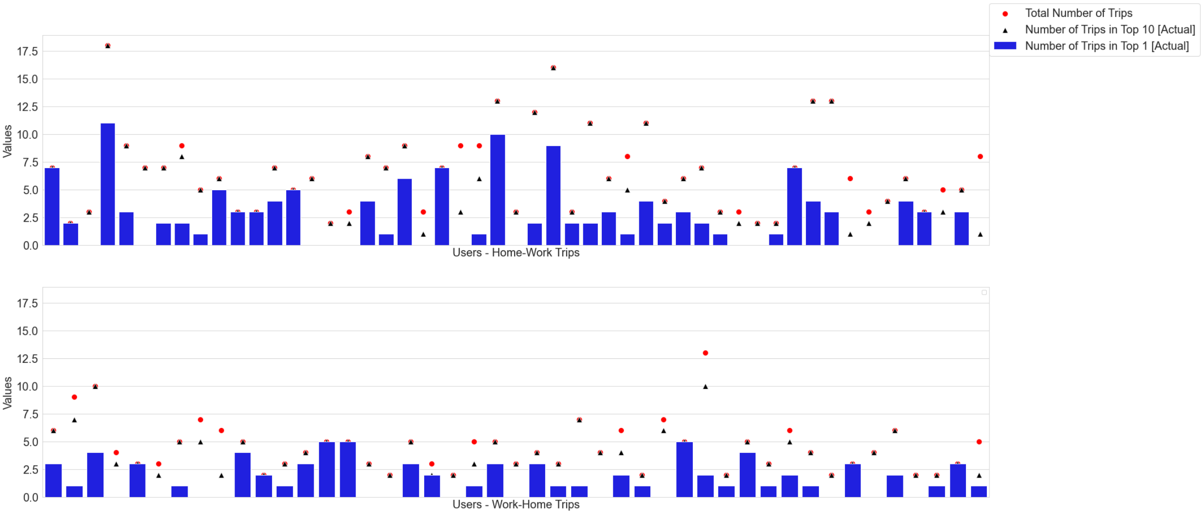
Starting with the actual path choice and timetable comparison, Fig. 10 shows, for each user in home-work trips (top) and work-home trips (bottom), the number of trips that matches the first position in the timetable CS (blue bar), the total number of trips (red circle), and the number of trips that are among the top 10 positions in the timetable CS.

Figure 10: Comparison of commuting trips path choices made by travellers with ordered timetable CS positions



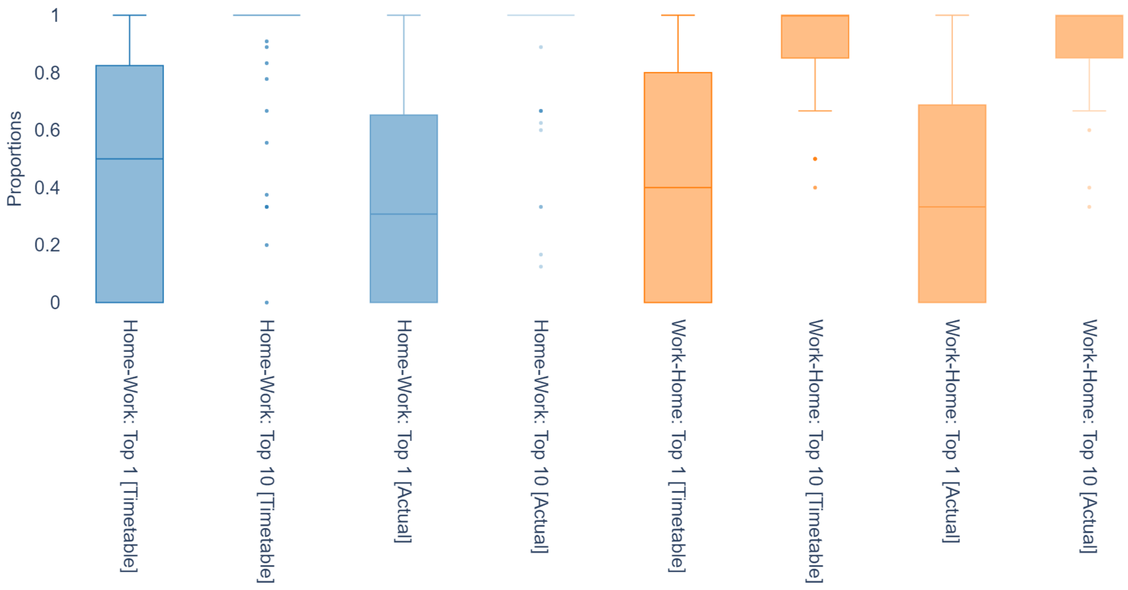
In Fig. 10, the proximity of the circle and triangle indicates the percentage of trips in that category that are ranked among the top 10 choices in terms of the shortest path plus a transfer penalty criterion. Moreover, if the blue bar is close to the circle, then most trips are optimal in terms of the criterion adopted. Fig. 11 repeats the same analysis for the actual CS.

Figure 11: Comparison of commuting trips path choices made by travellers with sorted actual CS positions



In order to compare the path choices of travellers in terms of actual and timetable CS, Fig. 12 shows the box plots of the proportions (for each traveller) of the number of trips in each category (top 1 and top 10, timetable and actual) over the total amount of trips made by each user. These proportions reflect the ratio of the blue bar over the red circle (top 1 proportion) and the ratio of the black triangle over the red circle (top 10 proportion).

Figure 12: Box Plots of proportions (for each traveller) of trips in the top 1 or in the top 10 of CS over total number of trips (home-work and work-home trips)



In Fig. 12 the first four box plots from left to right (in blue) correspond to home-work trips and the last four (in orange) to work-home trips. Moreover, the first two box plots in each trip category reflect the results for the timetable CS and the last two for the actual CS.

The first analysis is related to the differences observed between home-work and work-home trips. In general, home-work trips have a greater number of travellers whose proportions of top 1 and top 10 trips, relatively to the total of trips, exceeds 0.5 (first and third quartiles, as well as the median values, are all higher for home-work trips). In practice, the data reveals that home-work trips have a greater proportion of trips within the top 1 and top 10 positions of timetable and actual CS when compared to work-home trips.

The second analysis is the comparison between the proportions of trips when the timetable CS is contrasted with the actual CS. The differences are better observed by comparing the different box plots for the top 1 position. In this case, the proportions of top 1 trips in the timetable CS are higher than the proportions of top 1 trips in the actual CS, with higher medians and higher upper quartiles, suggesting that many users rely on the timetable to do their trips and stick to these choices even when the realized (actual) times for the trips are suboptimal from a duration and transfer penalty standpoint. The top 10 box plots, on the other hand, show identical lower quartiles (1 for home-work trips and 0.85 for work-home trips) for both timetable and actual trips, revealing that, even when the timetable option is not top 1 in the realized CS, it is most likely among the top 10, suggesting that only small disruptions, if any, are present. The information available to the travellers and the effects of disruptions are further discussed in the next subsection.

4.4 Impact of disruptions on commuting trips routes and information available to the traveller

Passive GPS tracking places a low burden on users, however, it also limits the amount of information that is available to the analyst. For example, even though online traffic information is made available to travellers by PT providers, it is unclear, by only analysing tracking data, whether the traveller has access to that information and the effects of that information on the route choice.

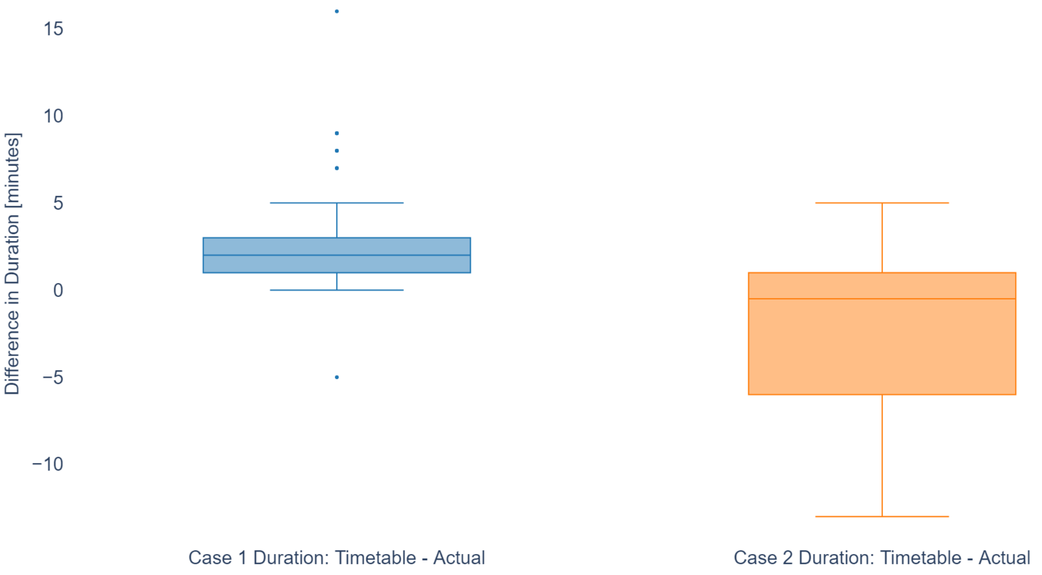
The existence of disruptions in the network may give some insights on which information was available to the PT user and the user's profile in terms of trip re-planning under disruptions. Of course, a major challenge is that there is not a clear distinction between what is a small and a big disruption, and the impacts of each on the user's end choice. In practice, disruptions are better defined in a continuous range of times, with different measurable impacts (Marra and Corman, 2020b). Even under this broader consideration, the impacts in terms of trip re-planning are highly subjective, as each traveller has their own tolerance to disruptions, which depends not only on a user profile but also on several externals factors that may continuously change, such as expectations on trip duration, reliability of the alternative option, etc.

In this paper, similar route choices were identified for commuting trips from home to work and vice-versa. For the subset of travellers in these trips, for which a default route choice is assumed, studying whether a disruption has caused a change in the typical route provides useful information regarding the traveller's features related to tolerance for trip re-planning as well as insights on which information was available to the traveller. In Zürich, most lines running inside the city operate under short headways, and big delays (over several minutes) are very rarely observed, unless there is a major accident or event (such as line closure, power outage, etc.). For the analyses that follow, it is further assumed that the route choices of travellers are not affected by capacity constraints and/or crowding, and that passengers may react to small disruptions by only changing their route and not by, for instance, opting for a private mode of transport or even changing their final destination. Furthermore, disruptions are considered with respect to total duration times. Hence, if, for the same route, there is a significant difference between the total duration of the timetable and the realized trip, then a disruption has happened to that route.

This subsection addresses two cases involving travellers with commuting trips that were grouped together according to the similarity criterion. For the first case, a subset of these trips ranked top 1 according to the timetable CS (best option in terms of total duration and number of transfers) was taken. Among those, the ones for which the same route had an inferior position ("suboptimal") in the ranking for the actual CS were further selected. The duration of the optimal (timetable) was compared to the duration of the actual route for each trip (seconds were ignored and all times were rounded to the closest integer in minutes). Most trips in this subset had a delay associated, and in only a few cases the time difference was equal to zero or negative, indicating that a few alternative routes had very similar times or small advancements relatively to the timetable schedules. A total of 65 trips from 30 different users were found, with a mean delay of 2.72 minutes and a standard deviation of 3.11 minutes. Only one case had a negative difference associated (the actual duration of the trip was better than the duration of the first option in the timetable CS). Most trips had a small delay associated, which made them shift from the top 1 position in the timetable CS to worse positions in terms of the actual CS. Even though they were "suboptimal", they were still the choice made by travellers, revealing that either the travellers had no information about the delay or that, despite having information on the delay, they still stuck to their original choice. Fig. 13 shows the box plot (on the left) of the differences in duration of these trips.

The second case comprises trips for which the opposite was observed, i.e. trips that were optimal (top 1) from the perspective of the actual CS and suboptimal (higher positions) in the timetable CS. Given that the routes represent common route choices for that traveller, this indicates that the traveller was able to identify a route for which the actual duration was faster than the timetable duration, and ended up being the optimal choice in terms of realized times. This could have happened for different reasons, for which two are highlighted. The first is that the traveller stuck with their common option, and short advancements on this route (or, equivalently, delays in the alternative routes) made it optimal in terms of realized travel times compared to the other options in the timetable. The second one is that the traveller had access to online information about possible disruptions/advancements, and chose the fastest route consciously. However, these trips were far more uncommon than the trips in the first case. A total of 28 trips from 22 different users were found, or an average of 1.27 trips per traveller. For these trips, the difference between timetable duration and actual duration had a negative mean of −2.04 minutes, so actual times were, on average, faster than timetable times. Times had also greater variability than in case one, with a standard deviation of 4.40 minutes. Fig. 13 shows the box plot (on the right) of the differences in duration of these trips.

Figure 13: Box Plots of the difference in time duration (in minutes) of timetable and actual routes



While taking a subset of similar commuting trips offers the possibility to reasonably infer that the observed route choice was rationally made by the traveller, it has the obvious drawback that only small disruptions/advancements are taken into account, since different route choices than the ones classified as similar are not included in the subset. The goal was to study whether these common route choices were optimal from the perspective of total travel times and number of transfers, as discussed in the previous subsection and, in addition, to know whether this subset of trips could be used to infer about travellers' access to information based on response to actual duration times. In order to do that, the path identified for each trip was contrasted with timetable and realized times for not only the path chosen by the user, but also by multiple other PT choices. The chosen path and all other choices were ranked based on a criterion involving total duration plus transfer penalty. In theory, if realized times were exactly the same as timetable times for all trips, no differences in the position on both CS should be observed. In practice, a difference in position implies some disruption/advancement in at least one of the trips. In this subsection, the focus was on trips classified as optimal in terms of either the timetable CS (case 1) or the actual CS (case 2). For case 1, the traveller's choice is ranked first in the timetable CS and a change in the position is observed in the actual CS, and in case 2, the opposite happens, with the traveller's choice being ranked first in the actual CS and in a different position in the timetable CS. In both cases 1 and 2, it can be inferred that travellers are concerned about travel times and have a clear preference for routes that minimize travel times, so studying the differences in the positions of timetable and actual CS of these trips provide means for checking which information was available to the users. For this subset of commuting trips, which included 60 different users and 514 trips (324 home-work and 190 work-home), 93 of them, or about 18%, fell into either case 1 or case 2 above. In particular, case 1 comprised 65 trips (about 13%) for which the travellers were likely following the timetable and had either no information about the disruption on the route or, despite having information on the disruption, they still stuck to their original choice. For the remaining 28 trips (about 5%) in case 2, the two most plausible explanations were that either the travellers stuck with their usual option, and short advancements on the route (or, equivalently, delays in the alternative routes) happened, or travellers had access to online information about possible disruptions/advancements, and chose the faster route consciously.

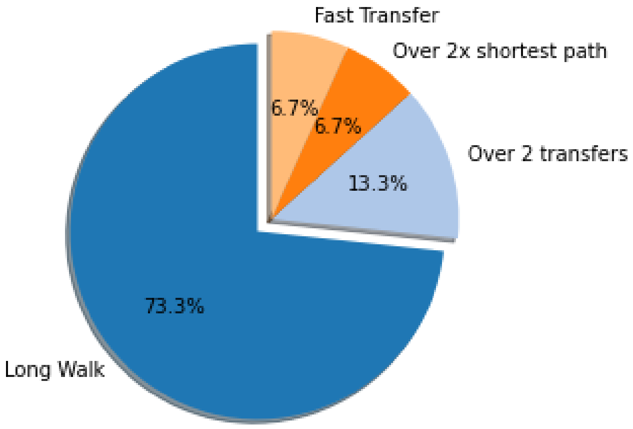
For the other trips not classified into cases 1 or 2, it is interesting to highlight that 180 trips (or about 35%) were ranked first in both timetable CS and actual CS, from which 118 were home-work and 62 were work-home trips. If the top 10 positions are considered, the number increases to 481 trips, or about 88%, where 290 were home-work and 161 were work-home. Hence, only about 12% of the commuting trips grouped by the similarity criterion had routes out of the top 10 positions in both timetable and actual CS.

A special case of the trips corresponds to the ones that are not contained in the CS. Although most travellers display similar behaviour when choosing a route on commuting trips, by prioritizing faster routes with a small number of connections, there are other attributes affecting the choice of some travellers. GPS tracking data allows only for limited knowledge of the travellers' utility function and perception of the network, based on an assumed model. For the computation of CS, only the relevant alternatives to passengers, according to the utility function considered, are included and ranked from 1 to 100. For the 590 commuting trips identified, the number of trips not included in either the timetable or the actual CS was 38 (6%), and the number of trips for which a trip was included in at least one of the two CS was 60 (10%). A total of 27 travellers had at least one commuting trip out of one CS, with an average of 2.22 trips out of

CS per commuter, and a standard deviation of 2.06 trips. However, the distribution of those numbers is right-skewed, with values up to the 59th percentile equal to 1, and only 11 travellers exceeding 1. In particular, the two travellers who had the highest number recorded (8 non-trivial trips each), had their trips not identified for two main reasons: the first traveller for making 3 transfers in 7 of 8 of his commuting trips, and the second for long walks (above the threshold) in all trips.

In general, for trips not contained in the CS, the traveller's choice was peculiar in the sense that it exceeded some logical conditions in the thresholds utilised for the choice set generation algorithm (Marra and Corman, 2020a), for example, the threshold for walking distance which was set to 750 m. Alternatively, the choice may simply have been poor from the perspective of the utility function utilised, causing it not to be ranked among the first 100 trips. A distinction must be made between the trips that were not identified in both actual and timetable CS and the trips that got identified in only one of them. While the first case may indicate that the traveller's choice was really poor in terms of minimizing total times, the second case may be related to other reasons. As examples, among all trips, two of them were ranked first in the timetable CS but did not show up in the actual CS, indicating that a major disruption may have happened. On the other hand, five trips were ranked first in the actual CS but did not show up in the timetable CS, which indicates that some advancements possibly made the trips (e.g. transfers) possible. Having these distinctions in mind, the main factors making these trips not being assigned to a CS are further investigated and depicted in Fig. 14.

Figure 14: Commuting trips outside of CS: depiction of factors.



A long walk appears as the factor having the highest impact in commuting trips not being listed on at least one of the CS, with 73.3% of the cases (or 44 trips), followed by the number of transfers exceeding the threshold (13.3% or 8 cases), and fast transfers (which cannot match the walking speed in the model) and too long duration both accounting for about 6.7% of the cases (4 trips each). If all trips are considered (and not only the commuting trips), the first factor remains in its position, but other factors are observed: 41% for a long walk; 28% for a too long duration; 14% of paths are after the 100th position; 7% are dominated paths; 4% do more transfers; 4% do a fast transfer, 2% used two times the same stop.

Two groups can be distinguished in terms of these trips: the first group with longer walks and longer duration paths, and the second group which aims for faster trips, for which travellers are either willing to make more than two transfers or are able to make faster transfers (e.g. walking pace above the assumed for the CS generation algorithm of 1.5 m/s). Even though the long walks associated with the commuting trips led to increased travel times, because most of them were isolated cases, it is not possible to infer any general behaviour. Only one traveller had, among their path choices, consistently picked the path with a longer walk, also increasing the travel time when compared to the shortest timetable/actual path, revealing a " willingness to walk" behaviour. If more cases like that were present, a CS generation algorithm that could identify this behaviour and adjust the thresholds for specific travellers would definitely have improved accuracy. For the investigation proposed in this paper, the number of commuting trips not listed in any CS is low, and most travellers seem to aim for a common goal (shortest path) and follow one of the top 10 trivial paths suggested in the timetable. However, in a big data scenario, simply discarding these trips implies losing a potential source of information that could enhance urban transport and urban mobility as a whole, not to mention market opportunities.

5 Conclusions and Future Research

This paper investigates longitudinal data on commuting behaviour in public transport based on travel diaries collected by a smartphone application called *ETH-IVT Travel Diary* consisting of 2901 PT trips of 172 users in the city of Zürich (Switzerland). The framework assumes unlabelled GPS tracking data, building up from the selection of an adequate unsupervised clustering technique to inferring the traveller's behaviour and information provision given a set of reasonable alternatives.

The proposed investigation on commuting behaviour is divided into two main parts. The first part considers all trips from home to work, and vice-versa, to find spatial and temporal regularities in terms of mode, line, departure times and trip duration. When applicable, the results are also contrasted with those of non-commuting trips. The second part introduces a heuristic algorithm to subset the commuting trips into sets of similar trips. The trips in these subsets, for each user, are assumed equivalent from a route perspective, and since they represent usual (and repeated) choices of users, their utility is compared with the utility of alternative routes generated by two choice sets, namely the timetable and the actual (realized) one. In most cases, the user option is equivalent to the optimal choice in terms of both choice sets. However, the interesting cases, which comprise routes with small disruptions and/or advancements, are utilised for inference on user's behaviour. More specifically, the trips for which the first timetable CS trip matched the usual user's path, but not the actual CS trip, allow inferring that the user stuck to the original plan (timetable), even though the trip was not optimal in terms of the realized times. Similarly, the trips for which the first actual CS trip matched the usual user's path, but not the timetable CS trip allow inferring that the user had access to online information and took the shortest route and/or unplanned advancements made the route optimal.

The findings in the paper can be summarized as follows: (1) commuters have regularity of departure times, and most trips are realized within 15 minutes of the average departure time (69% and 58% for home-work and work-home trips, respectively); (2) mode share of commuting and non-commuting trips differs substantially in terms of percentages of mixed-mode trips, especially those involving bus and tram (about 19.2% for commuting trips and only about 4.5% for non-commuting trips); (3) while commuters have preference for a single mode of transport (most of the trips were based on only one mode), line transfer is frequent, with 1.88 lines used per trip and the majority of trips (63%) having at least one line transfer; (4) although the network allows taking the same lines in both directions (home-work or work-home), only for about 50% of the commuters there was at least one trip with an agreement between the lines used in both directions; (5) important differences between the duration of home-work and work-home trips arise in transfer and walking times and, although the median values are similar, the analysis on the upper quartiles reveals a difference of 4 minutes (work-home trips taking longer), where two minutes come from walking times and about one minute from transfer times; (6) a significant part of the commuters are consistent in their chosen routes and departure times when those are considered simultaneously, more specifically, the proportion of similar trips among the home-work trips has a mean of 72%, while work-home proportion has a mean of 81%, so most commuters stick to a usual choice; (7) the majority of commuters choose a route that matches the optimal alternative (shortest path) of both timetable and actual CS, although the proportions of routes in the top 1 position of the timetable CS are higher than those of the actual CS; (8) disruptions/advancements were found in 18% of the trips studied under the similarity criterion, and in about 70% of them the commuter stuck to their usual (timetable) choice, although better choices were available according to the actual CS; (9) for the remaining 30% of these trips, travellers were likely to have online information about possible disruptions/advancements, and picked the optimal choice from the actual CS (equivalently, advancements on the route or delays in the alternative routes may have happened). In general, these results validated the consistency of commuters' behaviour and provided useful insights for policymakers to design and plan the PT network.

Many ideas for future work arise. First, the interpretation of commuting patterns by distinguishing trips by mobility scores based on spatiotemporal regularities, instead of using some heuristics to brute-force the subsets based on lines, departure times and duration. These mobility scores would allow for regression-based inference and, therefore, computing associated random errors. Furthermore, this research could be extended in the sense of looking not only at home and work OD pairs but also at all different OD pairs which represent regularly visited locations. This is especially relevant in times where home-office is consistently increasing as a result of more flexible work environments.

6 References

Bhadane, C. and K. Shah (2020) Clustering algorithms for spatial data mining, paper presented at the *Proceedings of the 2020 3rd International Conference on Geoinformatics and Data Analysis*, 5–9.

Bovy, P. H., S. Bekhor and C. G. Prato (2008) The factor of revisited path size: Alternative derivation, *Transportation Research Record*, **2076** (1) 132–140.

Clark, B., K. Chatterjee and S. Melia (2016) Changes to commute mode: The role of life events, spatial context and environmental attitude, *Transportation Research Part A: Policy and Practice*, **89**, 89–105.

Cottrill, C. D., F. C. Pereira, F. Zhao, I. F. Dias, H. B. Lim, M. E. Ben-Akiva and P. C. Zegras (2013) Future mobility survey: Experience in developing a smartphone-based travel survey in singapore, *Transportation Research Record*, **2354** (1) 59–67.

de Freitas, L. M., H. Becker, M. Zimmermann and K. W. Axhausen (2019) Modelling intermodal travel in switzerland: A recursive logit approach, *Transportation Research Part A: Policy and Practice*, **119**, 200–213.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu *et al.* (1996) A density-based algorithm for discovering clusters in large spatial databases with noise., paper presented at the *kdd*, vol. 96, 226–231.

Goulet-Langlois, G., H. N. Koutsopoulos and J. Zhao (2016) Inferring patterns in the multi-week activity sequences of public transport users, *Transportation Research Part C: Emerging Technologies*, **64**, 1–16.

Hong, R., W. Rao, D. Zhou, C. An, Z. Lu and J. Xia (2020) Commuting pattern recognition using a systematic cluster framework, *Sustainability*, **12** (5) 1764.

Kung, K. S., K. Greco, S. Sobolevsky and C. Ratti (2014) Exploring universal patterns in human home-work commuting from mobile phone data, *PloS one*, **9** (6) e96180.

Kusakabe, T. and Y. Asakura (2014) Behavioural data mining of transit smart card data: A data fusion approach, *Transportation Research Part C: Emerging Technologies*, **46**, 179–191.

Levinson, D. and S. Zhu (2013) A portfolio theory of route choice, *Transportation Research Part C: Emerging Technologies*, **35**, 232–243.

Lima, A., R. Stanojevic, D. Papagiannaki, P. Rodriguez and M. C. González (2016) Understanding individual routing behaviour, *Journal of The Royal Society Interface*, **13** (116) 20160021.

Liu, X., Q. Huang and S. Gao (2019) Exploring the uncertainty of activity zone detection using digital footprints with multi-scaled dbscan, *International Journal of Geographical Information Science*, **33** (6) 1196–1223.

Ma, X., C. Liu, H. Wen, Y. Wang and Y.-J. Wu (2017) Understanding commuting patterns using transit smart card data, *Journal of Transport Geography*, **58**, 135–145.

Ma, X., Y.-J. Wu, Y. Wang, F. Chen and J. Liu (2013) Mining smart card data for transit riders' travel patterns, *Transportation Research Part C: Emerging Technologies*, **36**, 1–12.

Marra, A. D. (2021) Tracking passengers to analyse travel behaviour during public transport disturbances, Ph.D. Thesis, ETH Zürich, Zürich, Switzerland.

Marra, A. D., H. Becker, K. W. Axhausen and F. Corman (2019) Developing a passive gps tracking system to study long-term travel behavior, *Transportation research part C: emerging technologies*, **104**, 348–368.

Marra, A. D. and F. Corman (2020a) Determining an efficient and precise choice set for public transport based on tracking data, *Transportation Research Part A: Policy and Practice*, **142**, 168–186.

Marra, A. D. and F. Corman (2020b) From delay to disruption: Impact of service degradation on public transport networks, *Transportation Research Record*, **2674** (10) 886–897.

Molloy, J., A. Castro Fernández, T. Götschi, B. Schoeman, C. Tchervenkov, U. Tomic, B. Hintermann and K. W. Axhausen (2020) A national-scale mobility pricing experiment using gps tracking and online surveys in switzerland: Response rates and survey method results, *Arbeitsberichte Verkehrs-und Raumplanung*, **1555**.

Molloy, J., C. Tchervenkov and K. W. Axhausen (2021) Estimating the external costs of travel on gps tracks, *Transportation Research Part D: Transport and Environment*, **95**, 102842.

Ortega-Tong, M. A. (2013) Classification of london's public transport users using smart card data, Ph.D. Thesis, Massachusetts Institute of Technology.

TomTom, T. I. (2021) Zurich traffic: Weekly traffic congestion by time of the day, `https://tomtom.com/en_gb/traffic-index`. Accessed: 14.01.2022.

Xiong, X., S. Qiao, N. Han, F. Xiong, Z. Bu, R.-H. Li, K. Yue and G. Yuan (2020) Where to go: An effective point-of-interest recommendation framework for heterogeneous social networks, *Neurocomputing*, **373**, 56–69.

Zhao, Y., X. Zhu, W. Guo, B. She, H. Yue and M. Li (2019) Exploring the weekly travel patterns of private vehicles using automatic vehicle identification data: A case study of wuhan, china, *Sustainability*, **11** (21) 6152.

Zhou, J., E. Murphy and Y. Long (2014) Commuting efficiency in the beijing metropolitan area: An exploration combining smartcard and travel survey data, *Journal of Transport Geography*, **41**, 175–183.

Zimmermann, M. and E. Frejinger (2020) A tutorial on recursive models for analyzing and predicting path choice behavior, *EURO Journal on Transportation and Logistics*, **9** (2) 100004.